Market Dynamics and Investment in the Electricity Sector*

Joseph A. Cullen[†] Stanley S. Reynolds[‡]

Revised, June 5, 2022

Abstract

A transition to a low carbon future will include a medium-to-long run period in which intermittent renewables co-exist with conventional fossil fuel electricity generators. Fossil fuel generators have frequent startups and shut-downs during the transition. We develop a dynamic competition model that allows for cycling of conventional generators. We apply the model to analyze long run effects of renewable investment subsidies and carbon prices in the Electric Reliability Council of Texas system. Reducing CO2 emissions via carbon pricing saves $165 million/year relative to a subsidy policy. We find significant differences in predictions between our model and a static model that ignores startup costs.

*We thank Ignacio Esponda, Gautam Gowrisankaran, Derek Lemoine, Glenn MacDonald, Ignacia Mercadal, seminar participants at Iowa State and the Berkeley Energy Institute, and conference participants at the AEA annual meeting, the Southern Economic Association annual meeting, and the Lancaster University conference on auctions, competition, regulation and public policy for helpful comments; and Chuan Chen for research assistance.

†Uber Technologies, Inc., jacullen@gmail.com

‡Department of Economics, Eller College of Management, University of Arizona, Tucson, AZ, sreynold@arizona.edu

1 Introduction

Production of electricity and heat is the largest source of greenhouse gas (GhG) emissions world-wide; see IPCC [2014]. Many countries and states around the world have implemented incentives for expansion of renewable energy generation from sources such as wind and solar, as a way of mitigating GhG emissions. Such policies, coupled with falling prices for renewable technologies, have spurred rapid growth in renewable electricity generation. Electricity markets are highly regulated and will likely experience increased government involvement to reshape the grid towards a cleaner production portfolio. In this paper we analyze the medium-to-long run economic effects of renewable energy incentives in a market in which solar and on-shore wind are important renewable energy sources. Understanding the economic effects of these incentives is critical for evaluation of policies that promote a transition to a low carbon electricity system. The economic effects we consider include impacts on energy prices, welfare costs of emissions reductions, and investment in electricity generation capacity.

An important aspect of this research is the role played by startup (or, cycling) costs for conventional fossil-fuel electricity generators. These generators are likely to remain as a significant part of the generation portfolio for several decades during a transition to a low or no-carbon future. As penetration of intermittent renewables increases, conventional generators must startup and shut-down more frequently to offset fluctuations in renewable generation; see Troy et al. [2010] and Perez-Arriaga and Batlle [2012]. Greater penetration of renewables and more frequent cycling of conventional generators may lead to significant changes in the volatility of wholesale energy prices, in profits of conventional generators, and ultimately to changes in investment incentives for various types of conventional generators. We seek to understand how the economic forces unleashed by a large scale renewable energy expansion play out in long run equilibrium.

We address our research question by first developing and analyzing a formal dynamic competitive equilibrium model that incorporates both short-run operating dynamics associated with electricity generator startup costs and long-run generator investment decisions.

We apply our analysis to the electricity system of the Electric Reliability Council of Texas (ERCOT). Data from ERCOT and other sources are used to parameterize our theoretical model. We utilize counterfactual simulations to examine the impact of two policy options: a renewable investment subsidy and a carbon tax. We compare outcomes under these policies for several CO2 emissions reduction targets. The counterfactual simulations permit us to identify the role that startup costs play in the response of investments, energy prices, and welfare cost of emissions reduction to renewable incentives.

What does our economic analysis contribute? While there is a large literature that examines the impact of renewable energy incentives, our analysis is unique in employing a dynamic market equilibrium framework to address the research question. This framework allows us to capture short-run operating dynamics and long-run investment incentives. Inclusion of conventional generator startup costs will alter the distribution of wholesale electricity prices, for example by raising prices during peak and/or shoulder hours and lowering prices during off-peak hours. These effects on prices change long run generator investment incentives and affect the mix of generators in the portfolio and the welfare cost of emissions reduction. The approach developed here may be used to analyze the impact of other renewable policies, such as feed-in tariffs and clean energy standards. Moreover, an economic model such as the one we develop here is useful for assessing the value of complementary policies, such as demand-response policies and energy storage incentives.

Our dynamic competition model allows for aggregate demand shocks, one-time capacity investment decisions from a menu of technologies, repeated generator startup/shut-down decisions, and production decisions for active firms. Firms incur a lump sum cost upon each generator startup, which introduces dynamic linkages across periods. Our model formulation presents theoretical and computational challenges, due to production non-convexities and the absence of an easily derived steady state equilibrium. We show that a dynamic competitive equilibrium is the solution of a social planner's problem. We characterize properties of competitive equilibrium and use the planner's problem as a computational platform for our application to electricity market analysis.

In the theoretical portion of the paper we prove equivalence between dynamic perfectly competitive allocations and the solution to a particular stochastic dynamic programming (DP) problem for a planner via a 'small firms' assumption. This assumption allows us to side-step equilibrium existence difficulties posed by production non-convexities, while retaining key economic implications of non-convexities. We use the equivalence result to show existence of a competitive equilibrium. We allow heterogeneity in types to emerge endogenously in equilibrium via firms' initial capacity investment decisions. All firms with a given type of technology are symmetric, but the number (or, measure) of firms of each type is endogenous. We further characterize properties of equilibrium allocations analytically using a special case of the model. Here we show how demand variability coupled with the magnitude of startup costs influence the extent of heterogeneity. Specifically, more frequent demand shifts and/or higher startup costs for baseload technology make it more likely that a higher cost, but more flexible, technology will be invested in.

The electricity system of the Electric Reliability Council of Texas (ERCOT) serves as a test bed for our analysis. ERCOT operates a self-contained grid that serves most of the state of Texas. Strong wind resources and favorable public policies have resulted in Texas having the highest wind turbine capacity of all U.S. states; LaRiviere and Lu [2020]. Texas has also experienced rapid growth in solar photovoltaic (PV) capacity in the last decade. We apply our theoretical model to analyze wholesale electricity market competition in ERCOT, focusing on the long-run impact of policies that promote investment in renewable energy. We show that incorporating startup costs for conventional natural gas fired generators yields more volatile electricity prices and a different mix of generator investments, compared to predictions from a conventional 'grid-stack' model of generator operations; we refer to this type of model as a static model.[1] These differences are magnified as renewable penetration increases. The presence of startup costs in the model changes equilibrium wholesale electricity prices in ways that reduce the profitability of renewable generators, implying that higher subsidies or higher carbon prices would be required to achieve CO2 reduction targets. We find that investment subsidies required to achieve a 60% CO2 reduction target are 14% higher for our

[1] We use the term 'static model' to refer to a model with no startup costs and no minimum output rate, and hence no dynamic linkages across operating periods.

dynamic model with startup costs than for the static model. A similar difference between dynamic and static model results arises for a carbon pricing policy. The presence of startup costs in the model has significant effects on equilibrium outcomes even though total startup costs represent a modest percentage of firms' overall costs.

The effectiveness of policies may be measured and compared by finding the reduction of total (consumer plus producer) surplus per ton of CO2 emissions reduction. We find that a carbon pricing policy is more effective than a renewable investment subsidy policy. For a 60% CO2 reduction target the average cost of CO2 emissions reduction is 17% higher under a renewable subsidy policy than under a carbon pricing policy. This yields a $165 million per year cost difference between the policies for the ERCOT system. While there is a substantial cost difference between these two policies, some other studies find larger differences in policy effectiveness. For example, using ERCOT data from 2008, Fell and Linn [2013] find that a renewable production subsidy or a renewable portfolio standard would be 4 - 6 times more costly than a carbon pricing policy. Our finding is likely driven by the fact that our long run equilibrium has no investment in coal-fired generators; thus, carbon pricing does not serve to induce coal-to-gas switching as it does in other studies. We also find that dynamic frictions affect estimates of policy effectiveness. Our estimates of welfare cost per ton of CO2 emissions reductions for high CO2 reduction are 60% higher in a dynamic model with startup costs than in a static model, for both types of renewable policies.

Our analysis builds on the dynamic competition literature; see Lucas and Prescott [1971], Jovanovic [1982], Rob [1991], and Hopenhayn [1990]. As in these papers, firms in our model are price-takers who face stochastically varying market conditions and make forward-looking decisions based on rational expectations of future outcomes. Our model departs from these papers in two economically important ways, however. First, we specify a fixed, lump sum cost each time a firm starts a generator for production. This formulation is in contrast to continuity assumptions for firms' state transitions used in other papers.[2] Our assumption of

[2] In Lucas and Prescott [1971] this takes the form of convex costs of adjustment of continuous state variables. Jovanovic [1982] and Hopenhayn [1990] assume continuously distributed costs of entry and exit, and/or probabilities of transitions between discrete states that are continuous in firms' actions. Continuity of state transition probabilities in players' actions is also a key assumption in Adlakha et al. [2015], which analyzes stationary equilibria of dynamic games with many players.

fixed, lump sum startup costs permits us to parameterize the model in the application without resorting to specifying distributions for startup costs. We prove existence of a dynamic competitive equilibrium for our model, in the absence of continuity for firms' state transitions. Second, many dynamic competition papers permit firms to make one-time entry and exit decisions. Our model of startups and shut-downs may be interpreted as a model in which individual firms are persistent, and repeatedly enter and exit the market. This is important in settings where firms are long lived and can wait for optimal startup (entry) and shut-down (exit) times. Unlike prior work, firms are ex-ante heterogeneous when startup (entry) and shut-down (exit) decisions are made. Combined with persistence, this permits us to model situations where differences between types are fixed and permanent once established, but at the same time allows the distribution of types which arises via equilibrium investments to be endogenous.

Several prior studies assess the long run impact of large scale renewable energy penetration. A common approach is to use a perfect competition model with a structure similar to our model; a short-run component - capturing market clearing prices and operation of generators under conditions of fluctuating demand - as well as a long-run component - capturing profit maximizing investment in generation capacity. Bushnell [2010] uses this type of model to examine the impact of of increasing wind penetration on fossil fuel generation investment and wholesale prices in U.S. Western Electricity Coordinating Council regions. In a similar vein, Green and Vasilakos [2011] analyze the impact of an increase in wind capacity on the mix of generating capacity, wholesale prices, and emissions in Great Britain. Fell and Linn [2013] analyze the impact of several different types of renewable support policies on long run wholesale market outcomes, and apply the analysis to ERCOT. Blanford et al. [2014] analyze the effects of a clean energy standard on generation investment in the U.S., using the US-REGEN model. Gowrisankaran et al. [2016] analyze the effects of a renewable energy mandate for electricity generation in Arizona. They focus on how increased solar photovoltaic penetration affects operating reserves, investment in conventional generators, and emissions. However, none of these papers incorporate dynamic generator operating constraints into a model with optimizing agents, as we do in this paper.[3] The static competition model that we report on and use for comparison purposes in the application embodies the main features of the models used in the electricity studies cited above.[4] Our comparisons of static and dynamic model results for the ERCOT application demonstrate key differences between our results and the kind of results that emerge from prior long run studies of electricity market investment.

We note two limitations of our analysis of electricity markets. First, we invoke a small-firms assumption to facilitate our dynamic competition analysis. This implies that firms are price-takers, so that supplier market power is not considered in our analysis. There is a large literature that analyzes the extent and effects of market power in restructured wholesale electricity markets; see for example Wolfram [1999], Borenstein et al. [2002], Hortacsu and Puller [2008]. The small-firms assumption smooths out aggregate production non-convexities, which allows us to characterize a competitive equilibrium with linear energy prices. If firms are not small then non-convexities may require modifications to the equilibrium formulation. For instance, O'Neill et al. [2005] show that efficient market-clearing prices may be obtained by complementing energy prices with additional prices for generator startups.[5] Second, we do not consider electricity transmission constraints in our analysis, and instead analyze a single pooled wholesale electricity market. Transmission constraints can limit energy flows between nodes at some hours and can also be a factor in siting decisions for renewable energy generation facilities. In some regions the most productive renewable energy resources are far from demand centers, so that new or expanded transmission lines are required to increase renewable energy penetration. While these limitations may be important for mechanism

[3]The US-REGEN model employed by Blanford et al. [2014] includes minimum generation rate constraints and allows for startup costs. However, the model uses ad-hoc decision rules for startup decisions, rather than optimization-based decision rules.

[4]These long run studies tackle some issues that we do not take up. For example, Bushnell [2010] considers both an energy-only market design and a design that includes a capacity market; Green and Vasilakos [2011] consider both a perfect competition model and an oligopoly model of short run operations; Gowrisankaran et al. [2016] incorporate a system operator's decisions about operating reserves into the short run component of their model.

[5]ERCOT and a number of other wholesale markets employ uplift (or make-whole) payments to compensate firms for startup and other fixed costs when energy prices alone do not provide sufficient revenue to cover costs; see FERC [2014].

design and system operations issues associated with market power and renewable energy, we believe that our framework is useful for assessing how supply frictions impact the long run impact of renewable energy expansion.

Finally, we contribute to the emerging literature on the effects of supply frictions in electricity markets. Cho and Meyn [2011] analyze a dynamic model of a wholesale electricity market and show that generator ramping constraints may lead to significant wholesale price volatility and sustained periods of prices that deviate from marginal generation cost. Reguant [2014] and Cullen [2015] estimate dynamic structural models of electric generator operations that include startup decisions and startup costs. Both papers find economically significant estimated generator startup costs. Cullen [2015] uses estimated generation and startup costs to simulate a dynamic competitive equilibrium, which is used for counterfactual policy analysis. The counterfactual analysis of Cullen [2015] applies to short-run competitive equilibrium with fixed generation capacity, in contrast to the long-run investment analysis of the present paper. Reguant [2014] shows that accounting for generator startup costs yields smaller estimates of price-cost markups during peak demand periods.

The rest of the paper proceeds as follows. In section 2 we characterize the model of competition. In section 3 we develop the connection between market equilibrium and the solution to the planner's problem. In section 4 we describe the data we use, explain how we set parameter values, and describe the computation approach. In section 5 we report results. The conclusion is in section 6.

2 A Competitive Model with Operating Dynamics

We formulate a dynamic competition model in which firms make decisions about investment, production startups and shut-downs, and production quantities over time. Activity is divided into two stages. Firms make decisions about investment in stage one. Firms have a variety of technologies from which to choose for their initial investment. We assume that investment is irreversible, so that capacity investments made in stage one remain in place for all of stage two. The discount factor between stages one and two is $\tilde{\delta} \in (0, 1]$. Stage two is comprised of an infinite sequence of periods, where in each period firms make decisions about startups and shutdowns, and active firms make production decisions. The per-period discount factor within stage two is, $\delta \in (0,1)$. Stage one decisions may be viewed as firms' long-run investment choices, for which profitability of investment is driven by product market competition in stage two.[6]

2.1 Demand

Demand varies across time periods within stage two according to the value of a demand shock (or, shift) vector, $\boldsymbol{\theta}$, which is assumed to follow a Markov process. There is an inverse market demand function, $P(Q,\boldsymbol{\theta})$, that is continuous and decreasing in total output Q. Define Θ as the set of all possible $\boldsymbol{\theta}$-values. $P(0,\boldsymbol{\theta})$ is assumed to have a finite upper bound for all $\boldsymbol{\theta} \in \Theta$. In addition, we define the gross benefit function B by:

$$B(Q,\boldsymbol{\theta}) \equiv \int_0^Q P(z,\boldsymbol{\theta})dz. \tag{1}$$

2.2 Production

In stage one firms may invest in one of J different production technologies. We normalize the amount of production capacity that a firm may invest in to one for each technology. We assume that all capacity units for any technology $j \in \{1,...,J\}$ have identical costs and characteristics. We use the following notation:

c_j = marginal cost of output for technology j

f_j = investment cost per unit of technology j capacity

s_j = startup cost per unit of technology j capacity

m_j = minimum output rate per unit of technology j capacity; $m_j \in [0,1]$

[6]Our focus in this paper is on once-and-for-all long-run equilibrium investment decisions. It is possible to extend the model to allow for repeated investment decisions over time. Such an extension would allow examination of phenomena such as boom-and-bust cycles of investment and industry shake-out. We believe that analogues of results in Propositions 3.1-3.3 could be derived for a repeated investment model. Computations for applications would be considerably more complex than those we provide in section 5.

Firms choose whether or not to invest in stage one; if a firm invests it also chooses which technology $j \in J$ to invest in. The total amount of investment in type-j capacity by all firms is k_j, and total cost of that investment is $f_j k_j$. Firms that invest in stage one compete over an infinite sequence of periods in stage two. Marginal cost of output is assumed to be constant for each technology type. A firm whose type-j capacity is inactive at the start of the current period may begin production in the same period by incurring a startup cost of s_j.[7] A firm with an active unit is restricted to operate between the minimum and maximum output rate for their technology; type j units must produce at a rate in the interval $[m_j, 1]$ per active unit of capacity. A firm with an active unit at the start of a period may shut down immediately. Firms may startup and shut-down repeatedly in stage two. There are four exogenous parameters - c_j, f_j, s_j, and m_j - for each technology $j \in J$.

Startup costs introduce non-convexities into the production technology, which in turn complicates market analysis and may lead to non-existence of competitive equilibrium. In order to pursue our objective of a competitive market analysis that incorporates these technology features, we introduce a formulation in which non-convexities are permitted at the firm level, but for which the aggregate production technology is convex. We assume that individual firms are small relative to the size of the market; specifically, firms are assumed to be of measure zero in the formal model.

In sections 4 and 5 we apply this model to analyze competition in a wholesale electricity market. However, the model is not limited to this application. For example, in the context of the airline industry, the capacity investment decision might the be number of aircraft to purchase while the startup cost would correspond to the cost of opening a new route with the stock of existing aircraft. In a labor context, the startup cost might be the fixed cost of hiring/firing workers while the investment cost would be the investment in capital used by workers (office space, machinery, etc) to create output. The framework may be applied to situations where there are both long-run capacity investments and potential entrants which are persistent and heterogenous.

[7] The model may be extended to allow for lags in startups and/or shut-downs. A one period lag in startup and/or shut-down is a straightforward extension. Lags of more than one period require an expansion of the state space for the planner's problem, which complicates computation of the model.

2.3 Market, Feasibility, & Equilibrium

The supply side of the market is comprised of a large number of small firms who operate as price takers. In stage one there is a mass $\overline{k}_j$ of type-j firms that may invest in production technology j. The mass of type-j technology firms that invests is $k_j \in [0, \overline{k}_j]$.

The production technology for a firm that invests may be described quite simply. In period t in stage two a firm was either active or inactive in the preceding period; $\omega_{t-1} = 1$ indicates active status last period and $\omega_{t-1} = 0$ indicates inactive status last period. The firm chooses its status $\omega_t \in \{0, 1\}$ at the start of each period t. If a type j firm switches from inactive last period to active this period then the firm incurs startup cost s_j. Production for a type-j active firm is constrained to be between the min and max rates for its technology type; a type-j firm's output in t is $\gamma_t \in [m_j\omega_t, \omega_t]$.

The following notation is used to describe the aggregate production technology. A vector $\boldsymbol{x}$ indicates the amount of each type of capacity that is active in a period; x_{jt} is equal to the mass of type-j firms that have $\omega_t = 1$. The vector $\boldsymbol{q}$ is the amount of output from the J types of firms, where, $q_{jt} \in [m_j x_{jt}, x_{jt}]$ for $j \in \{1, ..., J\}$. The aggregate production technology is defined via the following constraints:

$$\boldsymbol{k} \in K \equiv [0, \overline{k}_1] \times ... \times [0, \overline{k}_J] \tag{2}$$

$$(\boldsymbol{x}_t, \boldsymbol{q}_t) \in A(\boldsymbol{k}) \equiv \{(\boldsymbol{x}, \boldsymbol{q}) : x_j \in [0, k_j]; q_j \in [m_j x_j, x_j]; j \in \{1, .., J\}\} \tag{3}$$

Capacities must satisfy the investment constraint, given by (2).[8] Constraint (3) specifies that active capacity in period t may not exceed total capacity for any technology, and total output in t must be between the minimum and maximum output achievable for a given amount of active capacity in t for each technology.

An allocation is defined by a a vector $\boldsymbol{k}$ of capacities and a sequence for $(\boldsymbol{x}_t, \boldsymbol{q}_t)$. In general, the values for this sequence of vectors will depend on realizations of the demand shock process, since as we will show, firms' decisions will depend on realizations of this

[8] This constraint is imposed for technical reasons. It insures that capacities are bounded for the planner's problem.

process. Because of this, it is useful to describe an allocation as a stochastic process. It will be convenient to denote a history of realizations of demand shocks through time period t as, θ^t, and the set of all possible histories through t as, Θ^t.

DEFINITION: A *feasible allocation* is a stochastic process $\{\boldsymbol{k}, \boldsymbol{x}_t, \boldsymbol{q}_t\}_{t=0}^{\infty}$ that
(i) is measurable with respect to the set of possible histories of demand shocks, and
(ii) satisfies $\boldsymbol{k} \in K$ and for each realization of the process, $(\boldsymbol{x}_t, \boldsymbol{q}_t) \in A(\boldsymbol{k})$ for $t \geq 0$.

The set of feasible allocations is convex, since K is convex and $A(\boldsymbol{k})$ is convex for each $\boldsymbol{k} \in K$. Note that the set of feasible allocations for the market is convex, even though the production possibilities set for an individual firm is not convex. Measurability with respect to demand shock histories essentially means that there is an outcome for vector $(\boldsymbol{x}_t, \boldsymbol{q}_t)$ corresponding to each possible demand shock history, θ^t.

In each period t of stage two firms are assumed to observe the current price p_t and the history of demand shocks through t, θ^t. Firms are assumed to have rational expectations regarding future prices. Equilibrium prices for a given period depend on the history of demand shocks through that period. A price process $\{p_t\}$ is a stochastic process that is measurable with respect to the set of all possible histories of demand shocks. Given a price process $\{p_t\}$, the value of a type-j firm in period t is given by:

$$v_{jt}(\omega_{t-1}, \theta^t) = \sup_{\{\gamma,\omega\}} \mathrm{E}_t \sum_{\tau=0}^{\infty} \delta^{\tau} [\omega_{t+\tau}\gamma_{t+\tau}(p_{t+\tau} - c_j) - \max\{\omega_{t+\tau} - \omega_{t+\tau-1}, 0\} s_j] \qquad (4)$$

The suprenum in (4) is with respect to stochastic processes for decisions (γ_t, ω_t) that are measurable with respect to the set of possible demand shock histories. A policy for decisions regarding startups, shut-downs and output is *profit maximizing* for a firm with own initial state ω_{t-1} if it attains (4).

A type j firm will weigh the expected discounted payoff in stage two against the cost of investment. The value function for a type j firm in stage one is,

$$v_j^e = \max\{0, \tilde{\delta} E[v_{j0}(0, \boldsymbol{\theta}_0)] - f_j\} \tag{5}$$

where the expectation on the RHS of (5) is taken over initial stage two values of the demand shock. Note that a type j firm is indifferent between investing and not investing if the second term in brackets on the RHS of (5) is zero. A policy for a type j firm is profit maximizing *iff* it attains (4) in stage two and attains (5) in stage one.

DEFINITION: An allocation $\{\boldsymbol{k}, \boldsymbol{q}_t, \boldsymbol{x}_t\}$ together with a price process $\{p_t^*\}$ is a *market equilibrium* if:

(i) The allocation is feasible,

(ii) The allocation is consistent with profit maximizing policies for all firms, and

(iii) $p_t^* = P(\sum_{j=1}^{J} q_{jt}, \boldsymbol{\theta}_t)$ for all $t \geq 0$.

Condition (ii) states that all firms adopt policies that attain (4) and (5) when faced with price process $\{p_t^*\}$. Condition (iii) is a standard market clearing condition.

3 Competitive Market Equilibrium

The key to our results is demonstration of equivalence between a competitive equilibrium allocation and the solution to a planner's problem. We show that a solution to the planner's problem exists, and so the resulting allocation and price process constitute a market equilibrium. The equivalence of a competitive market equilibrium and a planner's solution is, of course, a fairly standard type of result and parallels results for dynamic market equilibrium models in Lucas and Prescott [1971], Jovanovic [1982], and Hopenhayn [1990]. However, our formulation differs from the papers cited above in an important way, and demonstrating this connection requires proof.[9]

[9] The model in Lucas and Prescott [1971] has a single type of representative firm, in contrast to the model in this paper which has several types corresponding to different technologies and firm-specific 'on/off' states. The model in Jovanovic [1982] has heterogeneous firms, but no aggregate shocks. The formulation in Hopenhayn [1990] allows for aggregate shocks and heterogeneous firms via a distribution of firm-specific states.

In our model there is a firm-specific binary state variable indicating each firm's prior operating status; active or inactive. A firm must incur a lump sum cost in order to transition from inactive to active status. Firm-specific transitions are not continuous in firms' control variables under this formulation. Results from Hopenhayn [1990] may not be applied to our model because that paper assumes a continuity condition on firm-specific state transitions.[10] We address the discontinuity in firm-specific state transitions that arises in our formulation by exploiting the 'small firms' (measure zero) assumption. While firm-specific state transitions are discontinuous, the transitions for the aggregate states that are relevant for the planner's problem are continuous, and this allows for a solution to the planner's problem that is equivalent to a competitive equilibrium allocation. To put things differently, binary states for firms coupled with lump sum transition costs pose an analytical difficulty in a model with large (positive measure) firms. In such a model, supply functions are not continuous in market prices and a competitive equilibrium need not exist. The small firms assumption side-steps this difficulty and also provides a way for us to link the solution to a planner's problem to competitive equilibrium allocations.[11]

3.1 The Planner's Problem

Before turning to equilibrium results we describe the planner's problem, which has a recursive structure. Let $\boldsymbol{x}'$ denote the vector of active capacities in the previous period. In any period $t \geq 0$ in stage two the planner has access to a vector $\boldsymbol{k}$ of total capacities for the J technologies and makes operating decisions in each period after observing $(\boldsymbol{x}', \boldsymbol{\theta}) \in X(\boldsymbol{k}) \times \Theta$, where $X(\boldsymbol{k}) \equiv \{\boldsymbol{x} : x_j \in [0, k_j]; j \in \{1, .., J\}\}$. $(\boldsymbol{x}', \boldsymbol{\theta})$ serves as a state vector in stage two for the

[10]Hopenhayn [1990] provides a general framework for competitive dynamics. This continuity assumption is critical for his Theorem 1, which demonstrates existence of a competitive equilibrium. Models consistent with this continuous transitions assumption include: Lucas and Prescott [1971] which has each firm's next period capital as a continuous function of its current capital and investment, and Ericson and Pakes [1995] in which each firm's state is in a finite set with state transition probabilities that are continuous functions of its control variables.

[11]The equivalence of an equilibrium allocation and the solution to a planner's problem is important for our analysis because it provides a way to prove existence of market equilibrium and because it allows us to use the planner's problem as a vehicle for computation.

planner.[12] Operating decisions are embodied in vector, $(\boldsymbol{q}, \boldsymbol{x})$, where $\boldsymbol{q}$ specifies production rates and $\boldsymbol{x}$ is the vector of active capacities for the current period. The values of $\boldsymbol{x}'$ and $\boldsymbol{x}$ together imply aggregate startup and shut-down decisions. The single period payoff, H, for the planner is total surplus for the period, which is equal to gross benefit less production cost and startup costs.[13]

$$H(\boldsymbol{q}, \boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\theta}) = B(\sum_j q_j, \boldsymbol{\theta}) - \sum_j [c_j q_j + s_j \max\{x_j - x'_j, 0\}] \tag{6}$$

H is bounded and is concave in $(\boldsymbol{q}, \boldsymbol{x})$ for each $(\boldsymbol{x}', \boldsymbol{\theta}) \in X(\boldsymbol{k}) \times \Theta$. Concavity follows from concavity of B in total output, linearity of production costs in output, and convexity of startup costs in $\boldsymbol{x}$. That H is bounded follows from our assumption that $P(0, \boldsymbol{\theta})$ has a finite upper bound for all $\boldsymbol{\theta}$ and from the capacity constraints that bound outputs for each technology.

The planner makes operating decisions in stage two to maximize expected discounted total surplus, where the single period return is H in equation (6). This can be described by a stationary stochastic dynamic programming problem with the following Bellman equation,

$$W(\boldsymbol{x}', \boldsymbol{k}, \boldsymbol{\theta}) = \max_{(\boldsymbol{x}, \boldsymbol{q}) \in A(\boldsymbol{k})} \{H(\boldsymbol{q}, \boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\theta}) + \delta E[W(\boldsymbol{x}, \boldsymbol{k}, \boldsymbol{\theta}^+) \mid \boldsymbol{\theta}]\} \tag{7}$$

where $\boldsymbol{\theta}^+$ is the next period demand shock. The value function $W(\cdot)$ may be used to define the stage one capacity choice problem for the planner.

$$\tilde{W} = \max_{\boldsymbol{k} \in \boldsymbol{K}} \{-\sum_{j=1}^{J} f_j k_j + \tilde{\delta} E[W(\boldsymbol{0}, \boldsymbol{k}, \boldsymbol{\theta}_0)]\} \tag{8}$$

The expectation on the RHS of (7) is taken in stage one over initial stage two values of the demand shock.

[12]If $t = 0$ then we require that $\boldsymbol{x}' = \boldsymbol{0}$. Active capacity in $t = 0$ requires startup in $t = 0$.

[13]Our definition of total surplus implicitly assumes that there are not startups and shut-downs for a single type of technology in the same period. In the proof of Proposition 3.1 we show that this property holds in the planner's solution.

3.2 Equilibrium Results

Proposition 3.1. *An allocation* $\{\boldsymbol{a}_t\} = \{\boldsymbol{k}, \boldsymbol{q}_t, \boldsymbol{x}_t\}$ *and price process* $\{p_t\}$ *constitute a market equilibrium iff the allocation solves the planner's problem of maximizing discounted expected total surplus.*

Proofs are in the appendix. The *if* part of the proof of Proposition 3.1 is constructed by first showing that any welfare maximizing allocation, along with the associated market clearing price process, maximizes aggregate market profits of firms, taking the price process as exogenous. The second step is to show that there is an assignment of operating policies to individual firms such that aggregate market profit maximization implies maximization of individual firms' profits. The *only if* part of the proof uses concavity of the planner's single period return H and convexity of the set of feasible allocations, to show that no alternative feasible allocation yields higher payoff to the planner than a market equilibrium allocation.

Proposition 3.2. *A market equilibrium exists.*

The proof proceeds by showing that a solution to the planner's problem exists. An optimal policy for a planner's solution generates a feasible allocation and a price process which, by Proposition 3.1 constitute a market equilibrium. Note that we have existence of a dynamic competitive equilibrium here in spite of technology non-convexity at the firm level. Our assumption of small (measure zero) firms effectively smooths out what would otherwise be discontinuities in supply functions.

Firms' startup and shut-down decisions are embodied in the $\boldsymbol{x}$ vector in each period. These decisions effectively determine a step supply curve for each period. The height of each step is equal to marginal cost for a particular technology type and the width of the step is the corresponding amount of 'on' capacity for that technology type. The market clearing price is determined by the intersection of the demand curve for the period (which is conditional on $\boldsymbol{\theta}$) and the step supply curve for the period. Therefore the market clearing price will either be equal to marginal cost for one of the technology types or be determined by the intersection of demand with one of the vertical segments of the step supply curve. Startup costs have only an indirect effect on market clearing prices via their effect on the position of the step supply curve.

Proposition 3.3. *Let $\{p_t\}$ and $\{p'_t\}$ be two equilibrium price processes. Then $\{p_t\} = \{p'_t\}$ almost everywhere.*

The proof of Proposition 3.3 is almost identical to that of Theorem 2 in Hopenhayn [1990], and therefore omitted. The idea of the proof is that if there were two distinct equilibrium price processes then there must be two distinct equilibrium allocations. Social welfare is equal at these two equilibrium allocations, since equilibrium allocations maximize the planner's objective. But if equilibrium price processes are distinct then the marginal social value of output differs across the two allocations for some histories of demand shocks. This would imply that a convex combination of the two allocations, which is feasible by convexity of the aggregate technology, would yield strictly higher social welfare.

Note that the allocations that generate the unique equilibrium price process need not be unique. We provide an example at the end of sub-section 3.3 with two equilibrium allocations with differing amounts of investment in technologies 1 and 2 that generate the same equilibrium price process. Moreover, even for a specific allocation of aggregate outputs, startups, and shut-downs that support the price process, there may exist alternative assignments of active/inactive status for individual firms of a particular type that are consistent with the same total amount of active capacity for that type.

Corollary 3.4. *If the investment constraints (2) are not binding then firms earn zero expected profit in equilibrium.*

The Corollary is a direct result of the defintion of payoffs in (4) and the definition of equilibrium. If $k_j < \bar{k}_j$ in equilibrium then there is a positive measure of type-j that choose to not invest. If type-j investing firms earn positive profits, then non-investing type-j firms are not maximizing profit since they could earn greater profit by investing.

3.3 A Special Case

We derive results below for a special case of our model in order to illustrate properties of equilibrium. We assume a linear demand function, two demand states, and two technologies. In addition we assume no lag in long run investment (i.e., $\tilde{\delta} = 1$) and minimum production rates equal to 100% ($m_1 = m_2 = 1$). While these assumptions are clearly restrictive, the qualitative nature of results for this special case are likely to carry over to more general specifications. In particular the assumption of a 100% minimum production rate is strong, but serves to simplify the model and to highlight the role of startup costs.

Inverse demand is given by, $P(Q, \theta) = \theta - Q$. We assume that the demand shock takes on either value θ_A or θ_B each period, with $\theta_A > \theta_B$. Demand shocks follow a Markov process over an infinite horizon in stage two with, $\text{Prob}[\theta_{t+1} = \theta_i \mid \theta_t = \theta_i] = \rho$ for $i \in \{A, B\}$. We assume that $1/2 < \rho < 1$ so that demand in successive periods is positively correlated. The long run probability of each demand state is 1/2 due to the symmetry of transition probabilities. We assume that the initial distribution of θ matches the long run probabilities.

We allow for two technologies, with $c_1 < c_2$ and $f_1 > f_2$. We define:

$$\Delta \equiv c_2 + 2(1-\delta)f_2 - c_1 - 2(1-\delta)f_1$$

and assume cost parameters are such that $\Delta > 0$. The assumption $\Delta > 0$ implies that technology 1 has a combined production and investment cost advantage over technology 2.

We first consider the case in which there is investment only in technology 1 in equilibrium. Proposition 3.5 below provides a sufficient condition for this case. Given just two demand states and a single technology, there is a steady state equilibrium in which prices alternate between a peak price, p_A, and an off-peak price, p_B, as follows:

$$p_A = c_1 + 2(1-\delta)f_1 + (1-\delta\rho)s_1 \tag{9}$$

$$p_B = c_1 - \delta(1-\rho)s_1 \tag{10}$$

When the startup cost parameter s_1 is zero, equations (9) and (10) yield the familiar peak and off-peak prices from the peak-load pricing literature; Williamson [1966]. The off-peak price is equal to marginal operating cost (c_1) and the peak price is equal to the marginal operating cost plus the capacity rental rate ($(1-\delta)f_1$) divided by the frequency of the high demand state (1/2).

As the startup cost parameter s_1 increases, the off-peak price falls and the peak price rises. In particular note that the off-peak price is less than marginal operating cost by the amount, $\delta(1-\rho)s_1$. This amount is equal to the discounted, expected savings in next period startup cost associated with keeping capacity active in the low demand state. Firms are willing to operate at a loss in low demand states in order to avoid the cost associated with shutting down and then re-starting when demand reaches a high state. The peak price rises with higher values of s_1 since a firm must be compensated with a higher peak price to be willing to incur higher startup costs for transitions between low and high demand states. The peak price p_A includes the term, $(1-\delta\rho)s_1$, which we define as an startup premium. When a firm starts capacity in state A, the expected revenue associated with the startup premium for that startup episode is, $\sum_{\tau=0}^{\infty} \delta^\tau[\rho^\tau(1-\delta\rho)s_1]$. This expression is equal to startup cost, s_1. The upshot is that the startup premium component of the peak price yields exactly enough expected revenue to cover the startup cost.

The amount of capacity started up during transitions between low and high demand states ($k_1^* - x_1''^*$) is increasing in the discount factor δ and demand persistence parameter ρ. While increases in these parameters reduce the peak price, they also raise the expected duration of a sequence of high demand states or yield less discounting of future payoffs for high demand states. The latter effects outweigh the price effects of greater demand persistence and higher discount factors. Note also that the amount of capacity started up during transitions between low and high demand states is decreasing in s_1.

Next we consider the case in which firms invest in both technologies in equilibrium. The structure of equilibrium mirrors the case of investment in a single technology. Output is equal to total capacity, $k_1 + k_2$, in all periods with high demand (state A). If the initial demand realization is low (state B) then startup capacity and output are equal to x_1'; at the first transition to state A there is startup equal to $k_1 - x_1'$ for technology 1 firms and k_2 for technology 2 firms, and output rises to $k_1 + k_2$. In any transition from state A to B all technology 2 firms shut down and output drops to k_1. In subsequent transitions from B to A all technology 2 capacity starts up. Technology 2 is the 'swing technology', utilized only during high demand states. In the appendix we derive the following solution for steady state equilibrium prices:

$$\tilde{p}_A = c_2 + (1 - \delta\rho)s_2 + 2(1 - \delta)f_2 \quad (11)$$

$$\tilde{p}_B = c_1 - \delta(1 - \rho)s_1 - \frac{1 - \rho\delta}{\delta(1 - \rho)}(\Delta - (1 - \rho\delta)(s_1 - s_2)) \quad (12)$$

In an equilibrium with technology 2 firms, steady state prices are similar to those for an equilibrium with only technology 1. The peak price $\tilde{p}_A$ has the same form as the peak price in (9), with technology 1 cost parameters replaced by technology 2 cost parameters. The off-peak price $\tilde{p}_B$ is equal to the off-peak price in (10) plus an additional term proportional to $\Delta - (1 - \rho\delta)(s_1 - s_2)$. This expression plays a key role in the following proposition.

Proposition 3.5. *If $\Delta > (1-\rho\delta)(s_1-s_2)$ then firms invest only in technology 1 in equilibrium and steady state equilibrium prices are defined in (9) and (10). If $\Delta < (1 - \rho\delta)(s_1 - s_2)$ then firms invest in both technologies in equilibrium and steady state equilibrium prices are defined in (11) and (12).*[14]

Proposition 3.5 provides conditions under which an startup cost advantage for technology 2 is large enough to offset its production and investment cost disadvantage. Let $s_2 = \beta s_1$, where β is a fraction less than one; a low value for β indicates a large startup cost advantage for technology 2. Then the second inequality in Proposition 3.5 may be expressed as:

$$\rho < (1 - \frac{\Delta}{s_1(1 - \beta)})/\delta \quad (13)$$

Figure 1 illustrates how a startup cost advantage for technology 2 and demand persistence

[14] By construction, the equilibria characterized in this proposition satisfy incentive compatibility conditions for firms. Consider for example, the case of $\Delta > (1-\rho\delta)(s_1-s_2)$, in which only type 1 firms invest and steady state prices are given by (9) and (10). Suppose that a type 2 firm were to invest, startup at each transition to state A, and shut-down at each transition to state B. The expected value of profits for this type 2 firm simplify to, $[p_A - c_2 - 2(1-\delta)f_2 - (1-\delta\rho)s_2]/(2(1-\delta))$. This expression is negative if $\Delta > (1-\rho\delta)(s_1-s_2)$.

influence equilibrium investment. Parameter values for (β, ρ) that lie in the shaded area in Figure 1 satisfy inequality (13). Such parameter values are consistent with a large startup cost advantage for technology 2 and relatively low demand persistence; these parameters induce equilibrium investment in both technologies. Parameter values that lie above the curve in Figure 1 induce equilibrium investment only in technology 1. Note that a high value of ρ requires a large startup cost advantage for technology 2 in order for firms to invest in technology 2 in equilibrium. Put somewhat differently, as ρ falls and startups and shut-downs become more frequent, a small startup cost advantage is enough for technology 2 to offset its production and investment cost disadvantage.

A knife-edge case with multiple equilibria arises if $\Delta = (1 - \rho\delta)(s_1 - s_2)$. Values of (β, ρ) that satisfy this knife-edge case lie on the curve that separates the shaded and unshaded areas in Figure 1. There is one equilibrium in which only technology 1 firms invest. Other equilibria involve investment in both technologies, with at least some of the startups that occur in transitions between states B and A coming from technology 2 firms. Note that steady state equilibrium prices defined in (9) - (10) and in (11) - (12) are the same for this knife-edge case, which is consistent with Proposition 3.3.

These results illustrate the importance of short-run dynamics and startup cost differences for equilibrium investment results. Under the $\Delta > 0$ assumption, the 'static' model with no startup costs implies no investment in the higher marginal cost technology. When positive and heterogenous startup costs are taken into account, the dynamic model may yield investment in both technologies. Moreover, the incentive for investment in the high marginal cost technology is sensitive to the persistence of demand shocks.

4 Application to Investment in Electricity Markets

We extend the model developed above to allow for intermittent renewable generation and then apply the extended model to analyze the effects of environmental policies on electricity market outcomes. Specifically we investigate the long-run effects of renewable energy incentives on a typical electricity market by calibrating the model and numerically solving for equilibrium outcomes for different levels of renewable investment subsidies and carbon prices. We compare our results to those from a static model (without startup costs) to examine how ignoring market frictions would misrepresent market outcomes.

This application maps well into the fundamentals of the model. First, electricity markets are characterized by large, long-lived investments in a variety of well-understood generating technologies. Once built, generator characteristics are more or less fixed and it is up to the owners of an asset to decide how best to utilize it. Second, many generating technologies face costs associated with participating in the market. To supply electricity to the market, a conventional generator must startup and begin operating at a sufficiently high level. Starting up generators is costly for operators both in terms of fuel and mechanical wear-and-tear; see Kumar et al. [2012]. Finally, demand in electricity markets is quite variable across seasons and even within a day. Demand volatility necessitates that at least some generators startup and shut down frequently. An increase in demand volatility will intensify the role of startup costs in determining market outcomes. Increasing penetration of intermittent renewable energy generators will tend to increase the uncertainty and volatility of the residual demand facing conventional generators.

To operationalize the model, we first construct operational and investment cost parameters for each technology type. Second, we characterize the stochastic processes governing demand shocks and renewable generation.

4.1 Technology Parameters

We calibrate the parameters for four types of electricity generation technology: natural gas combined cycle (GCC), natural gas turbine (Peaker), wind turbines, and solar PV.[15] GCC and Peaker units are conventional fossil-fuel generation technologies. Conventional generator operators have a high level of control over their production level, incur significant marginal costs of production and need to be forward looking due to production adjustment costs. Wind

[15] Investment in coal generation was included in the analysis for our 2017 working paper. In the intervening years, new investment in coal generators has become much less attractive in the U.S. due to falling natural gas prices and lower wind turbine and solar PV investment costs.

and solar are included as the renewable technologies. These technologies have no fuel costs, but their production depends on the availability of wind or solar energy. Their maximum production per unit of capacity in a given hour is referred to as the hourly capacity factor. The stochastic processes governing wind capacity factors and solar capacity factors are estimated based on ERCOT generation data. The solar PV data is from utility-scale installations. There is also distributed solar PV located at homes and business sites. We are not able to observe generation from distributed solar since this occurs behind the customer electricity meters. We allow for curtailment of wind and solar generation, so that these generators are able to produce less energy than the amount indicated by their capacity factor in a given hour.

The technology parameters can be split into two types. First, there are the characteristics of each technology that determine how an existing unit of capacity will compete in the daily operations of the market. These include fuel costs, emissions costs, minimum operating rate, and generator startup cost. Second, there are costs of investment that drive the decision to invest in a unit of capacity of that technology. These costs are based on data on construction costs, maintenance costs, and the expected lifetime of the generator. Costs are calibrated using data from the Energy Information Administration (EIA), the Electric Reliability Council of Texas (ERCOT), and estimates from the previous literature. The full details of our calibration approach may be found in Appendix B.

Generator parameters are shown in Table 1. The first row shows the range of marginal costs across the different technologies. The marginal costs for GCC and Peaker units are for the baseline case with no emissions costs. We adjust their marginal costs when carbon taxes are considered. If technologies differed only based on their marginal costs then we might expect to see only one technology in the market. However, the technologies vary across other dimensions as well. For example, wind and solar have negligible marginal costs, but the highest investment costs. In addition, their production cannot be controlled, except via curtailment. Peaker plants have the highest marginal costs, but are very flexible with no startup costs. They also have the lowest investment cost. It is these tradeoffs that lead to a mix of technologies in the market.

4.2 Electricity Demand and Renewable Generation

There are three stochastic components to our model: demand shocks (or, shifts), wind capacity factors, and solar PV capacity factors. We operationalize the state vector $\boldsymbol{\theta} \equiv (\psi, \nu, \sigma, h)$ to have four elements: current demand shock ψ, current wind capacity factor ν, current solar capacity factor σ, and hour-of-day h. We assume that $\boldsymbol{\theta}$ follows a Markov process. We use a linear inverse wholesale market demand function:

$$P(Q, \psi) = \psi - bQ. \tag{14}$$

The current demand state ψ is the demand shock in a given hour. The parameter b is a constant slope term across all demand shocks. In order to calibrate this demand function we need to specify the slope parameter b as well as the distribution of demand shocks.

To estimate the process for $\boldsymbol{\theta}$, we use hourly data on load (quantity demanded), wind generation, solar generation, and wind and solar capacity levels from the Texas ERCOT market. Specifically, we estimate three separate regression equations for load, wind capacity factor, and solar capacity factor. The first equation regresses current hour load on hour-of-day dummy variables, lagged load interacted with hour dummies, squared lagged load interacted with hour dummies, lagged wind capacity factor interacted with hour dummies, and lagged solar capacity factor interacted with hour dummies. The second equation regresses current hour wind capacity factor on the same set of explanatory variables as the first equation, except that lagged squared load is replaced with lagged squared wind capacity factor. The third equation regresses current hour solar capacity factor on the same set of explanatory variables as the first equation, except that lagged squared load is replaced with lagged squared solar capacity factor. This estimation is explained in more detail in appendix B. Though we use data from the Texas grid, the observed data patterns are similar to those in other parts of the southern and western U.S.

To calibrate the slope parameter b, we look to the literature on electricity demand. Short-run wholesale electricity demand is very price inelastic, although the elasticity is probably not zero as is sometimes assumed in analyses of wholesale electricity markets. While most retail electricity customers in the U.S. are subject to relatively inflexible retail prices, many commercial and industrial customers are exposed to wholesale price fluctuations. Commercial and industrial customers in ERCOT may bilaterally negotiate power contracts with retailers, and these contracts have a variety of provisions, including time and seasonally-varying rates, critical peak reductions, transmission charges, and real-time-pricing; see Xie et al. [2013]. The direct and indirect responses of large customers to wholesale price changes lead to some price-responsiveness of wholesale demand. Price elasticity estimates for market level wholesale electricity demand vary from zero to -0.35, with results concentrated in the low end of that range; Patrick and Wolak [2001], Johnsen [2001], Xie et al. [2013], Bigerna and Polinori [2014]. We choose a value for price elasticity of -0.1, and set the value of parameter b to yield this price elasticity at the average price and quantity observed in the data.

We do not incorporate a long-run demand response to price changes. While electricity demand is more price elastic in the long-run than the short-run, recent empirical estimates of long-run elasticity are also fairly small in magnitude, in the range of -0.3 to -0.4; Deryugina et al. [2020]. We therefore expect that a long-run demand response to price changes induced by renewable energy policies would have second order effects on our results.

4.3 Computation

After calibrating parameters, we solve for competitive equilibrium investments in different technologies, generator outputs, and wholesale energy market prices. The solution method for the static model uses a straightforward nested optimization approach. Stage 2 involves maximization of total surplus for one year's worth of realizations of demand quantities, wind capacity factors, and solar PV capacity factors conditional on a vector of generator capacities. Stage 1 of the static model involves optimal choice of generator capacity levels to maximize surplus net of (one-year's worth of) capacity investment costs.

For the dynamic model, stage 2 of the planner's problem is an infinite horizon stochastic DP problem with a solution that is conditional on a vector of generation capacities. The state vector for this DP problem is $(x, \boldsymbol{\theta}) = (x, \psi, \nu, \sigma, h)$. State components x, ψ, ν, and

σ are continuous variables. The vector $\boldsymbol{\theta}$ constitutes the information set for constructing beliefs about demand shocks and renewable capacity factors in future periods. The model for forecasting future demand and renewable capacity factors is described in the previous sub-section.

We solve the stage 2 DP problem by utilizing the nonlinear certainty equivalent (NLCEQ) method described in Cai et al. [2017]. This method approximates the optimal policy function (i.e., optimal decision rule) at a finite collection of state values by solving deterministic finite horizon optimization problems in which future values of random state variables are set equal to their expected values conditional on the current state. This method can accommodate large DP problems with continuous state variables and control variables with inequality constraints. We use the computed policy function to calculate one year's worth of total surplus at a sequence of optimal x values and realized values of demand and renewable capacity factors.[16] The total surplus payoff for stage 2 is used as input to solve for optimal stage 1 capacity investments. By Proposition 3.1 the surplus maximizing allocation corresponds to a competitive equilibrium allocation.

The approach outlined above allows us to compute long run equilibrium solutions for the static model and the dynamic model for fixed policy parameters (an investment subsidy or a carbon tax). For the environmental policy analysis that follows we embed the problem of finding a long run equilibrium solution in a search problem with the objective of choosing a policy parameter (e.g., carbon tax) that yields a long run equilibrium that achieves a particular level of CO2 emissions. Additional details on the solution method may be found in appendix C.

5 Environmental Policy Analysis

We use the competitive model developed in section 3 coupled with model parameters specified in section 4 to assess the effects of two environmental policies - a renewable investment subsidy

[16]Note that the NLCEQ solution differs from a perfect foresight solution. Realizations of demand and renewable capacity factors will differ from forecasted values. The policy function arising from the NLCEQ method will involve adjusting decisions over time as realized values of random variables deviate from forecasts.

and a carbon tax - on electricity market outcomes. In order to facilitate policy comparisons we examine policy options that yield a fixed set of CO2 reduction target levels.

Almost all generation capacity investment in recent years in the U.S. has been in wind turbines, solar PV, and natural gas-fired units; see EIA [2020b]. From 2015-2019 over half of new U.S. generation capacity investment was in renewables, with about two-thirds of renewable investment in wind turbine capacity. Renewable mandates and subsidies have been critical for investment in wind. Federal production subsidies for wind and investment tax credits for solar have been renewed regularly, though the magnitude of federal incentives has been declining in recent years; see EIA [2021].

The short-run implications of increased wind power production have been studied extensively; see Cullen [2013], Kaffine et al. [2013], Fell and Kaffine [2014], Novan [2015], and Chyong et al. [2020]. Analyses of long-run effects of renewable energy mandates and incentives on generation investment and grid operations appear in Bushnell [2010], Green and Vasilakos [2011], Fell and Linn [2013], Blanford et al. [2014], and Gowrisankaran et al. [2016]. However, as we note in Section 1, these papers do not incorporate dynamic generator operating constraints into a model with optimizing agents, as we do in this paper.

Critics of renewable energy have argued that uncertain and volatile patterns of wind and solar electricity generation require expensive storage solutions or costly standby backup capacity. Land-based wind farms have highly variable production and tend to supply energy when it is least needed. Figure 2 shows the hourly average wind and solar capacity factors, along with electricity load (demand quantity), using 2019 data from ERCOT. As is clear from the figure, wind is most productive during hours of low demand. Solar generation tends to line up better with demand, but falls off prior to high demand in late afternoon and evening. Monthly average wind and solar generation and load exhibits similar patterns, again based on 2019 ERCOT data. Figure 3 shows that wind generation is lowest in the summer, when electricity demand is highest. Average monthly solar generation tracks average monthly loads fairly well; both peak in late summer months.

There is considerable hour-to-hour variation in wind and solar generation. This variation is important as it affects the variability of residual demand (load less renewable generation) served by conventional fossil fuel generators. We use observed wind and solar hourly capacity factors from 2019 data to project hourly renewable generation as renewable capacity increases. Figure 4 illustrates how increases in renewable generation capacity affect the average and standard deviation of hourly residual demand. For this figure we assume that the ratio of wind to solar generation capacity is 2:1, this is roughly the ratio that emerges in the policy simulations described below. Average residual demand falls steadily as renewable capacity grows, falling by 58% at the highest level of renewable capacity that we consider. The standard deviation of residual demand falls slightly as renewable capacity increases from a low level, but then rises, increasing by 32% for the highest renewable capacity we consider. Since conventional generators face production adjustment costs, this increased volatility in residual demand will impose real costs on the system and will favor flexibility over low operating cost. In the long run, this will change the portfolio of technologies in the market.

For the results that follow, we first solve the model without renewable incentives. In this baseline scenario, we compare the outcomes in the dynamic competitive equilibrium model with startup costs to outcomes in the static competitive model calibrated with the same parameters, except for startup costs. After this we report results for exogenous target levels of CO2 reduction that are achieved by renewable investment subsidies and by carbon taxes. This permits us to compare how results are affected by incentives under the two policies. In particular, we compare policy effectiveness, as measured by welfare cost per ton of CO2 reduction, for the two policies. In addition, for each policy setting we compare static and dynamic model predictions across several outcome measures. The results show that the presence of startup costs changes the distribution of equilibrium prices, the mix of technologies invested in, and how the technologies are utilized. Differences between static and dynamic model predictions increase as the level of renewable penetration rises.

5.1 Baseline Case - No Policy Incentives

Long run equilibrium results for the static model (no startup costs) and the dynamic model (with startup costs) are summarized in Table 2. Average long run equilibrium wholesale electricity prices are three percent higher for the dynamic model than for the static model. Substantial differences in investment between the models emerge, in spite of the models generating similar price levels. Most investment in the static model is in GCC capacity, with modest amounts of Peaker (natural gas turbine) and solar PV capacity. Wind turbines are slightly too costly to be profitable without policy incentives in the static model. The dynamic model yields less investment in GCC capacity and Peaker capacity, more investment in solar PV capacity, and positive investment in wind turbines.

Capacity investment differences appear to be driven by differences in patterns of wholesale prices between dynamic and static models. We compute the standard deviation of simulated wholesale price for each day of the year. Table 2 reports within day price standard deviations averaged across all days of the year. This measure of price dispersion is 15% higher for the dynamic model than the static model. In particular, equilibrium prices in late afternoon and evening hours (5 pm - 10 pm) are 10% higher in the dynamic model than in the static model. The presence of startup costs in the dynamic model means flexible Peaker units with high marginal cost are more likely to be turned on in late afternoon hours, with less reliance on costly short-term startups of GCC units. This pattern of price differences raises the profitability of wind turbine investment in the dynamic model. Renewable penetration is higher for the dynamic model than the static model with no renewable policy incentives.

There is no investment in coal-fired generators in either the static or dynamic equilibrium solutions. Given fuel price forecasts and new generator investment costs (EIA [2020a]), new investment in coal units is not profitable in ERCOT, even in the absence of carbon taxes or renewable subsidies.[17]

The presence of startup costs for GCC units in the dynamic model drives significant differences in renewable penetration and generator capacity investment between dynamic and static models. This is true even though startup costs incurred for GCC units are small; less than one percent of operating costs for GCC units in the baseline case.

[17] An earlier working paper, Cullen and Reynolds [2017], uses older data on investment costs and fuel costs and finds significant investment in coal-fired generators and no investment in solar PV capacity. The difference in results between static and dynamic models is more pronounced when coal-fired generators are part of the mix, because coal-fired units have high startup costs.

5.2 Policy Comparisons for the Dynamic Model

Next we evaluate and compare the long run impact of two policies: a subsidy for investment in renewable energy generation capacity and a carbon tax. We examine how these policies affect wholesale electricity prices, welfare, emissions, and generation investment in long run equilibrium for the dynamic model. We consider policy parameters to achieve three different target levels of CO2 reduction relative to CO2 emissions in the baseline case: 20%, 40%, and 60%. CO2 emissions for our baseline case are about 25% less than actual ERCOT CO2 emissions in our reference year of 2019, so these emission reduction targets would achieve even larger percentage reductions relative to 2019 actual emissions.[18]

For each type of policy we find the policy parameters that achieve target CO2 reductions. These policy parameters are reported in Table 3. Given the recent reductions in capacity investment costs for wind turbines and solar PV panels, only relatively modest investment subsidies or carbon prices are required to induce low or intermediate levels of CO2 reduction. However, to advance from 40% CO2 reduction to 60% reduction requires roughly doubling either the investment subsidy or the carbon price. A 32% subsidy is required to achieve a 60% CO2 reduction target under the subsidy policy, whereas a $31.60/ton tax on CO2 emissions is required to achieve the 60% reduction target under the CO2 tax policy. We frame the subsidies as investment tax credit (ITC) percentages, since the ITC has been commonly used as a renewable energy policy incentive by the U.S. federal government.[19]

Investment in wind turbines and solar PV capacity rise as renewable policy incentives increase. As renewable investment increases overall investment in conventional fossil fuel generators falls. However, the distribution of investment in GCC and Peaker generation units is quite different under the two policies. Figures 5 and 6 show changes in equilibrium capac-

[18] Renewable penetration for our baseline case is about half of renewable penetration for ERCOT in 2019. However in 2019 there was still significant generation from coal and this accounts for the higher CO2 emissions in the 2019 data.

[19] The ITC is commonly taken for solar PV projects. The current federal ITC for solar projects is 22 - 26%, depending on project dates. The federal production tax credit has been an important incentive for many wind turbine projects. The American Recovery and Reinvestment Act of 2009 allowed wind projects to take an investment tax credit in lieu of a production tax credit. Allowable ITC rates for wind projects have varied in recent years. The federal ITC for on-shore wind projects is 18% for projects begun before the end of 2021.

ities of GCC and Peaker units for progressively higher CO2 reduction targets. Investment in Peaker units increases, almost doubling from the baseline case to 60% CO2 reduction case under the subsidy policy. The dynamic model captures the rising value of flexible Peaker units in complementing increasing levels of intermittent renewable energy generation. Investment in Peaker units is roughly constant for different CO2 abatement targets under the carbon price policy. The rising value of flexible Peaker units with higher renewable penetration is offset by an increasing marginal operating cost disadvantage of Peaker units relative to GCC units as the carbon price rises.

A concern regarding grid integration of renewables is that high renewable penetration will depress wholesale prices to the point where investment in and electricity production from fossil fuel units would not be economic. This is the so-called merit order effect; see Sensfuß et al. [2008]. Introducing large amounts of renewable generation capacity with zero marginal operating cost could lead to many hours in which the wholesale electricity price is driven to zero. Table 3 summarizes equilibrium results across different levels of CO2 abatement. We find that increasing levels of renewable penetration driven by renewable investment subsidies tend to push down average long run equilibrium wholesale prices. Profitable investment in GCC and Peaker units continues to occur and indeed investment in Peaker units rises with higher renewable penetration under renewable investment subsidies. Average long run equilibrium wholesale prices rise with higher carbon prices. Investment in conventional fossil fuel generators falls with higher carbon prices, but wholesale prices net of carbon taxes are still high enough to support investment in conventional generators. Within day wholesale price variability rises with higher CO2 reduction targets under both policies. The standard deviation of hourly wholesale prices rises 50 - 80% as we move from the baseline case to 60% CO2 reduction.

Renewable policies affect the intra-day pattern of wholesale prices, as well as overall average wholesale prices. Wholesale prices do not fall uniformly for all hours as renewable penetration rises under a renewable subsidy policy. As we move from the baseline case to 40% CO2 abatement, average prices *rise* for early morning and evening hours and fall in night-time and mid-day hours, as depicted in Figure 7. These price effects mirror wholesale price impacts of increasing solar PV penetration reported by Bushnell and Novan [2018] in the California market. Their estimates imply that higher solar PV penetration substantially lowers wholesale prices mid-day but raises wholesale prices for 6 - 8 am and 7 - 9 pm. They highlight the difference between GCC units and more flexible Peaker units as contributing to these intra-day price impacts of solar. "Leading up to the morning ramp up in solar generation, and at the tail end of the evening ramp down in solar production, there is a shift away from more fuel efficient GCC production and towards less fuel efficient, higher marginal cost GT [Peaker] production."[20] Figure 8 provides the corresponding comparison of hourly wholesale prices for the baseline case and 40% CO2 abatement under carbon pricing. The intra-day pattern of wholesale prices for 40% CO2 abatement is similar for carbon pricing and renewable investment subsidy policies. The largest increases in wholesale prices under carbon pricing are for early morning hours and evening hours.

Increasing wind penetration displaces generation from conventional power plants and results in lower CO2 emissions. Table 3 reports policy effectiveness comparisons for the two policies across different CO2 reduction targets. We measure policy effectiveness by the loss in total surplus per ton of CO2 reduction relative to the baseline case; that is, average CO2 abatement cost. Note that investment subsidies (or foregone tax revenues) are not factored into welfare cost; these subsidies are viewed as transfers. The carbon tax policy is more effective than the subsidy policy, as we would expect. The policy effectiveness gap is small for a low CO2 reduction target but increases for higher reduction targets, reaching a 17% difference in average abatement cost for 60% CO2 reduction. Compared to a renewable investment subsidy policy, a carbon tax policy does two things to encourage efficiency. A carbon tax raises wholesale electricity prices and thereby reduces energy consumption and associated emissions. For a 60% CO2 reduction target the wholesale electricity price is nearly 40% higher under a carbon tax than under a subsidy policy. Given our parameterization

[20]Bushnell and Novan [2018], p. 16. Their results are for increasing solar penetration with a fixed set of conventional generators using data from CAISO, whereas our results are for equilibrium simulations of ERCOT with higher penetration for both wind and solar and allowance for long run conventional capacity adjustments. However, a similar mechanism appears to be at work in both settings. Relatively flexible Peaker generators are used rather than GCC generators during the morning solar ramp up and during the evening solar ramp down. This avoids some startup costs for GCC units that would otherwise be incurred.

of demand, this yields about a 4% reduction in energy consumption.[21] A carbon tax also incentivizes more efficient investment in and utilization of fossil fuel generation units, shifting investment and production away from Peaker units toward GCC units. These effects of a carbon tax policy allow CO2 reduction targets to be achieved with lower renewable energy penetration compared to a renewable investment subsidy policy.

Our results show that policy effectiveness of a carbon price is greater than policy effectiveness for a renewable investment subsidy. For a 60% CO2 reduction target the average cost of CO2 emissions reduction is 17% higher under a renewable subsidy policy than under a carbon pricing policy. This amounts to a $165 million per year cost difference between the policies for the ERCOT system. Some other studies find much lower policy effectiveness for renewable mandates and subsidies than for carbon taxes. For example, Fell and Linn [2013] examine policy effectiveness of carbon taxes and several renewable incentives schemes using a long run investment analysis based on ERCOT data. They find that welfare cost per ton of CO2 reduction for a renewable portfolio standard (RPS) is about four times greater than the corresponding cost for a carbon tax that yields the same wind penetration.[22] Several things might account for the difference in policy effectiveness comparisons between our findings and those of Fell and Linn [2013]. Fell and Linn [2013] use older data on investment costs and fuel prices and assume that existing coal generation plants continue to operate. Carbon prices play an important role in fuel switching between coal and natural gas when coal plants are a significant part of the generation portfolio. There is no investment in coal plants in our long run analysis and carbon prices play a more limited role, causing adjustments in relative amounts of investment and production in GCC and Peaker units.[23]

[21] Our estimate of the policy effectiveness gap is likely a lower bound since we do not consider long run demand responses. Adding a more elastic long run demand would cause a larger decline in energy consumption under a carbon tax policy.

[22] A RPS policy typically mandates a minimum percentage of electricity generation from renewable sources. A RPS could be implemented via investment subsidies, but a more common approach for state level RPS programs is implementation via establishment of renewable energy certificates, which provide additional payments to renewable energy generators for each unit of generation; see Fell and Linn [2013]

[23] A further qualification to our results on policy effectiveness is in regard to public financing costs. An investment subsidy (or, ITC) policy requires tax-financed government subsidies and a corresponding additional social cost of taxation. In contrast, a carbon pricing policy raises tax revenue and so reduces the overall taxation burden, leading to a reduction in social costs of taxation. This difference in public financing costs reduces the policy effectiveness of renewable investment subsidies relative to carbon prices.

5.3 Policy Comparisons: Dynamic v. Static Models

We now compare static model and dynamic model predictions of the effects of a renewable investment subsidy policy and a carbon price policy. To facilitate comparisons with dynamic model results we start with policy parameters for the static model that yield CO2 emissions equal to emissions from the no-policy baseline case of the dynamic model. Under the static model this level of emissions is achieved either with an ITC of 1.5% or a carbon tax of $1.40/ton.

Table 4 reports static and dynamic model predictions for renewable investment subsidies. The static model predicts that lower incentives are required to achieve CO2 reduction targets compared to dynamic model predictions. The required ITC for the static model is 3 - 6 percentage points lower compared to the dynamic model.

Average wholesale prices from static and dynamic models are similar. Within day price variability is higher for the dynamic model than for the static model. The gap widens with greater CO2 abatement. This difference in price variability is driven in part by differences in the pattern of average hour-of-day prices. The static model predicts higher night-time and mid-day prices, but lower prices for early morning hours and evening hours compared to the dynamic model.

The static model also yields different generation capacity investments compared to the dynamic model under the renewable subsidy policy. Static and dynamic model equilibrium capacity investments for GCC and Peaker units are depicted in Figures 9 and 10, respectively. The dynamic model predicts less GCC capacity investment and more Peaker capacity investment than the static model. For the highest level of CO2 reduction, the dynamic model predicts more than twice as much Peaker investment as the static model.

A key difference between static and dynamic models is their policy effectiveness predictions. The dynamic model predicts that CO2 reduction for the renewable investment policy is 70 - 95% more costly compared to static model predictions. Why is policy effectiveness lower in the dynamic model than in the static model? Several forces are at work. On shore wind turbines are most productive at night and solar PV is most productive mid-day. Equilibrium energy prices (and the marginal value of energy) tend to be lower for these hours in the dynamic model. This is reflected in greater curtailment of renewable energy in the dynamic model compared to the static model. As renewable penetration increases, GCC units have more frequent startups and shut-downs, leading to higher startup costs in the dynamic model (see Table 3). More frequent cycling of GCC units with higher renewable penetration changes investment, shifting the plant mix toward Peaker units with high operating costs and away from GCC units in the dynamic model. Peaker units have higher NOx and SO2 emissions per MWh of electricity generation than GCC units, so there is a significant environmental impact predicted by dynamic model that is not captured by static model.

The overall pattern of differences between static and dynamic model results for carbon pricing is similar to differences in results for renewable subsidies. Higher carbon prices are required with a dynamic model to achieve CO2 reduction targets. Carbon prices must be $2 - $4 per ton higher with a dynamic model to achieve targets compared to a static model. The policy effectiveness comparisons predicted by static and dynamic models for carbon pricing are similar to comparisons predicted for renewable subsidies. The dynamic model predicts that CO2 reduction for the carbon pricing policy is 70 - 95% more costly compared to static model predictions.

6 Conclusion

A transition to a low carbon future requires introduction of large amounts of intermittent renewable generation into the electricity system. There will likely be a significant period of time in which high penetration of intermittent renewable generators will co-exist with conventional fossil fuel generators. High renewable penetration means that conventional generators will have to startup and shut down frequently. More frequently cycling of conventional generators induces a number of changes in wholesale electricity markets.

We develop a model of dynamic competition that includes initial long run capital investments as well as short-run operating dynamics that allow for supply frictions due to power plant cycling. In our dynamic competition model firms that operate conventional power plants make an initial capital investment followed by repeated startup, shut-down, and re-start decisions as demand fluctuates. We extend prior theoretical research to show a correspondence between the model's dynamic competitive equilibrium and the solution to a social planner's problem. This allows us to establish equilibrium existence and also provides a computational platform for numerical analysis. A special case of the model with two demand states and two technologies is used to derive results on the impact of supply frictions and demand persistence on price variability and the distribution of technology investments.

We apply the framework to analyze the impact of renewable investment subsidies and carbon prices on long run outcomes, using data from the Texas ERCOT system. We incorporate startup costs of conventional generators as a supply friction in the market. The framework provides a tractable method of modeling competitive investment while accounting for operating dynamics introduced by startup costs. We find that incorporating generator startup costs into the model yields different electricity prices and long-run generator investments, compared to outcomes for a 'static' model without startup costs. This is the case even when incurred startup costs in equilibrium are relatively small. A key finding is that the presence of startup costs results in changes in the pattern of hourly wholesale electricity prices relative to predictions from a static model for a high CO2 reduction target. These wholesale price changes reduce the profitability of wind turbines and solar PV farms, so that a significantly higher investment subsidy (or carbon price) would be required to induce a target CO2 reduction level, compared to the subsidy (or carbon price) required under a static model. Similarly, the welfare costs of achieving emissions reductions under either investment subsidies or carbon pricing are estimated to be much greater for the dynamic model than for the static model. These results underscore the importance of incorporating short-run supply frictions when predicting long-run outcomes.

Our results should not be interpreted as an exhaustive policy analysis of integration of renewable energy into the electricity grid. There are many aspects of electricity markets that have been not been explored. For example, in order to highlight the economic forces in the model we maintain a parsimonious specification for technologies. We don't include technologies such as hydro power or nuclear that could be important for the future portfolio of technologies. We also excluded continued operation of legacy fossil fuel plants, such as coal-fired plants, in our policy analysis. One could also consider adding production flexibility into the model via large-scale energy storage. Our dynamic competition model may provide a good framework for analyzing energy storage issues as optimal storage decisions depend on forecasts of future wholesale electricity prices. These extensions are left for future research.

Appendices

A Proofs of Propositions

Proof of Proposition 3.1

We introduce a variation of the model that was specified in Section 2. Let y_{jt} be the amount of shut-down of type-j capacity in period t; $\boldsymbol{y}_t$ is the vector of these amounts for all technologies. Let z_{jt} be the amount of type-j capacity that starts in period t; $\boldsymbol{z}_t$ is the vector of these startup amounts for all technologies. Using this notation, the aggregate production technology may be modified so that equation (3) is replaced by:

$$y_{jt} \in [0, x_{j,t-1}], \forall j, \forall t, \tag{15}$$

$$z_{jt} \in [0, k_j - x_{j,t-1}], \forall j, \forall t, \tag{16}$$

$$x_{j,t} = x_{j,t-1} - y_{jt} + z_{jt}, \forall j, \forall t. \tag{17}$$

The set of feasible allocations is convex since the constraint sets are compact and equation (17) is linear. Also, using this notation, the planner's single period total surplus function may be defined as:

$$\tilde{H}(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) = B(\sum_j q_j, \boldsymbol{\theta}) - \sum_j [c_j q_j + s_j z_j] \tag{18}$$

$\tilde{H}$ is bounded and is concave and differentiable in $(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z})$. Note that we cannot have $y_{jt} > 0$ and $z_{jt} > 0$ in a socially optimal allocation for any $j \in \{1, 2, ..., J\}$. That is, an optimal allocation will not have simultaneous startup and shut-down for a single type of technology. This implies that maximal surplus for the planner's problem in the model of Section 2 and maximal surplus for the planner's problem in this variation of the model are identical. The version of the model with controls $\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z}$ has the advantage of differentiability of the planner's payoff, and this is the version used in the proofs.

(⇒) Let feasible allocation $\{\boldsymbol{a}_t\} = \{\boldsymbol{k}, \boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t\}$ be a solution to the planner's problem. This allocation is a stochastic process that is measurable on the set of possible histories of demand shocks. This allocation induces a process $\{\boldsymbol{x}_t\}$ for active capacities and a price process, $\{p_t\}$, where $p_t = P(\sum_j q_{jt}, \theta_t)$. Firms are atomistic price-takers and therefore view the price process as exogenous.

Let $\{\boldsymbol{a}'_t\} \equiv \{\boldsymbol{k}', \boldsymbol{q}'_t, \boldsymbol{y}'_t, \boldsymbol{z}'_t\}$ be some alternative feasible allocation. Define a family of allocations by, $\{\boldsymbol{a}^\lambda_t\} = \{\lambda \boldsymbol{a}'_t + (1-\lambda)\boldsymbol{a}_t\}$, for $\lambda \in [0,1]$. $\{a^\lambda_t\}$ is a feasible allocation since the set of feasible allocations is convex.

Since $\{\boldsymbol{a}_t\}$ is a solution to the planner's problem, we have,

$$-\sum_j f_j k^\lambda_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \tilde{H}(\boldsymbol{q}^\lambda_t, \boldsymbol{y}^\lambda_t, \boldsymbol{z}^\lambda_t, \boldsymbol{\theta}_t)] - [-\sum_j f_j k_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)]] \leq 0$$

for $\lambda \in [0,1]$. This implies that,

$$-\sum_j \frac{f_j(k^\lambda_j - k_j)}{\lambda} + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t [\frac{\tilde{H}(\boldsymbol{q}^\lambda_t, \boldsymbol{y}^\lambda_t, \boldsymbol{z}^\lambda_t, \boldsymbol{\theta}_t) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)}{\lambda}]] \leq 0$$

for $\lambda \in (0,1)$, and

$$\lim_{\lambda \downarrow 0}[-\sum_j \frac{f_j(k^\lambda_j - k_j)}{\lambda} + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t [\frac{\tilde{H}(\boldsymbol{q}^\lambda_t, \boldsymbol{y}^\lambda_t, \boldsymbol{z}^\lambda_t, \boldsymbol{\theta}_t) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)}{\lambda}]]] \leq 0. \qquad (19)$$

Gross benefit B is bounded above, since $P(0, \boldsymbol{\theta})$ has a finite upper bound for all $\boldsymbol{\theta}$ and quantities are limited by capacities. Therefore, H is bounded above. Since H is bounded above by a δ-integrable function, we can pass the limit operator inside the expectation and t-summation operator. So,

$$\lim_{\lambda \downarrow 0} -\sum_j \frac{f_j(k^\lambda_j - k_j)}{\lambda} + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \lim_{\lambda \downarrow 0}\{\frac{\tilde{H}(\boldsymbol{q}^\lambda_t, \boldsymbol{y}^\lambda_t, \boldsymbol{z}^\lambda_t, \boldsymbol{\theta}_t) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)}{\lambda}\}] \leq 0. \quad (20)$$

Taking the limits as λ approaches zero in (20) yields:

$$-\sum_j f_j(k'_j - k_j) + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \sum_j [(\frac{\partial B(\sum_j q_{jt}, \boldsymbol{\theta}_t)}{\partial Q} - c_j)(q'_{jt} - q_{jt}) - s_j(z'_{jt} - z_{jt})]] \leq 0$$

Define the price process $\{p_t\}$ to equal the derivative of the gross benefit function with respect to total output for each period for each $\boldsymbol{\theta}_t$. Then,

$$\begin{aligned} &-\textstyle\sum_j f_j k'_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \sum_j [(p_t - c_j)q'_{jt} - s_j z'_{jt}]] \\ &\leq -\textstyle\sum_j f_j k_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \sum_j [(p_t - c_j)q_{jt} - s_j z_{jt}]] \end{aligned} \tag{21}$$

The RHS of inequality (21) is aggregate market profit at the allocation $\{\boldsymbol{a}_t\}$ and price process $\{p_t\}$ induced by the solution to the planner's problem. Since $\{\boldsymbol{a}'_t\}$ is an arbitrary alternative feasible allocation, this inequality implies that, given $\{p_t\}$ and rational expectations on this process, there is no other feasible allocation that yields greater aggregate market profit than the allocation induced by the planner's solution.

The last part of the *if* proof is to connect aggregate market profit maximization to maximization of individual firms' profits. Given the allocation $\{\boldsymbol{a}_t\}$ we induce operating policies for firms as follows. For type-j firms, we assign the fraction $k_j/\overline{k}_j$ to invest and the remaining fraction to not invest. For period t and history θ^t there is a vector $(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{x}_{t-1})$ associated with $\{a_t\}$. For type-j firms, mass $x_{j,t-1}$ have $\omega_{t-1} = 1$ and mass $k_j - x_{j,t-1}$ have $\omega_{t-1} = 0$. If $x_{j,t-1} > 0$ then we assign the fraction $y_{jt}/x_{j,t-1}$ of active firms to shut-down in period t. If $x_{jt,t-1} < k_j$ then we assign the fraction $z_{jt}/(k_j - x_{j,t-1})$ of inactive firms in $t-1$ to startup and the remaining fraction of firms to remain inactive. All type-j firms that are active in period t are assigned output equal to $\gamma_t = q_{jt}/x_{jt}$. Any such assignment of operating policies yields a stochastic process for (γ_t, ω_t) that satisfies (4) for each possible demand shock history, for each firm. The assigned operating policy for a firm yields a (expected, discounted) profit for the firm, and total profits of all firms from period zero are equal to aggregate market profit on the RHS of (21). This equivalence holds since the operating policy assignment used implies that the aggregate production quantities and startup and shut-down capacities of firms add up to the q_{jt} quantities, z_{jt} startup capacities, and y_{jt} shut-down capacities on the

RHS of (21).

Suppose that there is an alternative policy for a firm that yields greater profit than the policy assigned based on allocation $\{\boldsymbol{a}_t\}$. Furthermore, suppose that the measure of firms for which this is true is positive. Then an alternative market allocation can be constructed such that the firms that achieve greater profit with an alternative policy are assigned the alternative policy, and other firms retain the policy based on allocation $\{\boldsymbol{a}_t\}$. This alternative market allocation is feasible, since there are no externalities in production across firms. The aggregate profit for this alternative market allocation exceeds aggregate profit associated with $\{\boldsymbol{a}_t\}$ (the profit on the RHS of (21)). But this contradicts the result that aggregate profits are maximized using allocation $\{\boldsymbol{a}_t\}$, given the exogenous price process $\{p_t\}$. That is, there cannot be a positive measure of firms that achieve greater profit than the profit associated with the assigned policy from allocation $\{\boldsymbol{a}_t\}$. If there is an alternative policy for a firm that yields greater profit than the policy assigned based on allocation $\{\boldsymbol{a}_t\}$, and the measure of firms for which this is true is zero, then the policy assignment for these firms can be changed to the alternative policy without altering the market allocation or the price process.

We have shown that socially optimal allocation $\{\boldsymbol{a}_t\}$, along with associated process $\{p_t\}$, satisfies the 3 conditions for a market equilibrium:
(i) The allocation is feasible, since a socially optimal allocation must be feasible,
(ii) There is an assignment of policies to firms that maximize profit for each individual firm,
(iii) The price process satisfies the market clearing condition, since prices are set to clear the market for each period t and history θ^t.

($\Leftarrow$) Let $\{\boldsymbol{a}_t\} = \{\boldsymbol{k}, \boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t\}$ and $\{p_t\}$ be the allocation and price process, respectively, for a market equilibrium. Suppose there is an alternative feasible allocation $\{\boldsymbol{a}'_t\} \equiv \{\boldsymbol{k}', \boldsymbol{q}'_t, \boldsymbol{y}'_t, \boldsymbol{z}'_t, \boldsymbol{x}'_t\}$ that is not a market equilibrium and yields greater total surplus than $\{\boldsymbol{a}_t\}$. Define $\{\boldsymbol{a}^\lambda_t\}$ as in the *if* part of the proof for $\lambda \in [0, 1]$; we know that $\{\boldsymbol{a}^\lambda_t\}$ is a feasible allocation. By the concavity and differentiability properties of $\tilde{H}$ in $(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z})$, using Lemmas 1 and 2 from Hopenhayn [1990], the function $(\tilde{H}(\boldsymbol{q}^\lambda_t, \boldsymbol{y}^\lambda_t, \boldsymbol{z}^\lambda_t, \boldsymbol{\theta}) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}))/\lambda$ is decreasing in λ, and increases to $\sum_j[(p_t - c_j)(q'_{jt} - q_{jt}) - s_j(z'_{jt} - z_{jt}))]$ as $\lambda \downarrow 0$. In addition, concavity of $\tilde{H}$ implies that $(\tilde{H}(\boldsymbol{q}_t^\lambda, \boldsymbol{y}_t^\lambda, \boldsymbol{z}_t^\lambda, \boldsymbol{\theta}) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}))/\lambda$ is bounded below by $(\tilde{H}(\boldsymbol{q}'_t, \boldsymbol{y}'_t, \boldsymbol{z}'_t, \boldsymbol{\theta}) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}))/\lambda$ for $\lambda \in (0, 1]$. These results are used in the string of inequalities below.

$$-\sum_j f_j k'_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \tilde{H}(\boldsymbol{q}'_t, \boldsymbol{y}'_t, \boldsymbol{z}'_t, \boldsymbol{\theta})] - [-\sum_j f_j k_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)]] > 0,$$

since $\{\boldsymbol{a}'_t\}$ yields greater total surplus than $\{\boldsymbol{a}_t\}$. This implies that,

$$-\sum_j \frac{f_j(k_j^\lambda - k_j)}{\lambda} + \tilde{\lambda}\mathrm{E}_0[\sum_t \delta^t [\frac{\tilde{H}(\boldsymbol{q}_t^\lambda, \boldsymbol{y}_t^\lambda, \boldsymbol{z}_t^\lambda, \boldsymbol{\theta}_t) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)}{\lambda}]] > 0$$

for $\lambda \in (0, 1]$, and

$$-\sum_j f_j(k'_j - k_j) + \lim_{\lambda\downarrow 0}\tilde{\delta}\mathrm{E}_0[\sum_t \delta^t [\frac{\tilde{H}(\boldsymbol{q}_t^\lambda, \boldsymbol{y}_t^\lambda, \boldsymbol{z}_t^\lambda, \boldsymbol{\theta}_t) - \tilde{H}(\boldsymbol{q}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{\theta}_t)}{\lambda}]] > 0.$$

This implies that,

$$\begin{aligned} &-\textstyle\sum_j f_j k'_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \sum_j[(p_t - c_j)q'_{jt} - s_j z'_{jt}]] \\ &> -\textstyle\sum_j f_j k_j + \tilde{\delta}\mathrm{E}_0[\sum_t \delta^t \sum_j[(p_t - c_j)q_{jt} - s_j z_{jt}]]. \end{aligned}$$

In other words, we have shown that the supposition that there is some alternative feasible allocation $\{\boldsymbol{a}'_t\}$ that yields greater total surplus than a market equilibrium allocation $\{\boldsymbol{a}_t\}$ implies that total market profits at $\{\boldsymbol{a}'_t\}$ exceed those at $\{\boldsymbol{a}_t\}$. This is inconsistent with the property that individual firm profits are maximized at $\{\boldsymbol{a}_t\}$, since the strict inequality above implies that at least some firms could earn greater profits by choosing a different policy.

Proof of Proposition 3.2

The planner's problem has two parts: the stage one choice of the vector $\boldsymbol{k}$ of capacities and the stage two choice of a policy function for output, startup and shut-down in each period of stage two, conditional on $\boldsymbol{k}$. By the usual backward recursion logic, consider stage two first. Recall that function $\tilde{H}$ is bounded and is concave, and therefore continuous, in $(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z})$ for each $\boldsymbol{\theta}$. Let $\boldsymbol{x}'$ be the vector of 'on' capacities from the previous period and denote the constraint set for the planner's choice of $(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z})$ in a period by $G(\boldsymbol{x}', \boldsymbol{k})$:

$$G(\boldsymbol{x}', \boldsymbol{k}) \equiv \{(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z}) : y_j \in [0, x'_j]; z_j \in [0, k_j - x'_j]; q_j \in [m_j(x'_j - y_j + z_j), x'_j - y_j + z_j]; j \in \{1, ..., J\}\} \tag{22}$$

G is compact valued for each $(\boldsymbol{x}', \boldsymbol{k})$ and continuous in $\boldsymbol{x}'$ and $\boldsymbol{k}$. Define an operator T on function u as follows:

$$(Tu)(\boldsymbol{x}', \boldsymbol{k}, \boldsymbol{\theta}) = \max_{(\boldsymbol{q},\boldsymbol{y},\boldsymbol{z}) \in G(\boldsymbol{x}',\boldsymbol{k})} \{\tilde{H}(\boldsymbol{q}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) + \delta \mathrm{E}[u(\boldsymbol{x}' - \boldsymbol{y} + \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{\theta}^+) \mid \boldsymbol{\theta}]\} \tag{23}$$

Let C be the space of bounded functions that are continuous in $(\boldsymbol{x}', \boldsymbol{k})$ that map $X(\boldsymbol{k}) \times K \times \Theta$ into $\mathbb{R}$. Then by an argument similar to that in the proof of Theorem 4.6 in Stokey and Lucas [1989], the operator T maps elements in C into C and T has a unique fixed point, which is the stage two value function $W(\boldsymbol{x}', \boldsymbol{k}, \boldsymbol{\theta})$ that satisfies the Bellman equation (7) for the planner.

It can be shown that the stage two value function $W(\cdot)$ is concave in $\boldsymbol{k}$. There exists an optimal choice of $\boldsymbol{k} \in K$ in stage one for the planner, since the payoff function is concave and the constraint set K is compact. The optimal $\boldsymbol{k}$ and the optimal policy associated with (7) yield a feasible allocation and a price process, which by Proposition 3.1 constitute a market equilibrium.

Proof of Proposition 3.5

Part One: Assume that, $\Delta - (1 - \rho\delta)(s_1 - s_2) > 0$. We suppose initially that firms invest only in technology 1 in equilibrium, and then show that this must hold. Given just two demand states and a single technology, the structure of equilibrium is fairly simple. Output is equal to total capacity k_1 in all periods with high demand (state A). If the initial demand realization is low (state B) then startups and output are equal to x'_1; at the first transition to state A there is startup equal to $k_1 - x_1'$ and output rises to k_1. In any transition from state A to B there is shut-down equal to $k_1 - x_1''$ and output drops to x_1''. In subsequent transitions from B to A there is startup equal to $k_1 - x_1''$. The payoff $\tilde{W}$ to the planner may be expressed as a function of (x_1', x_1'', k_1).

$$\begin{aligned}\tilde{W}(x_1', x_1'', k_1) = & \tfrac{1}{2(1-\delta)}[B(k_1, \theta_A) - c_1 k_1] + \tfrac{1}{2(1-\delta\rho)}[B(x_1', \theta_B) - c_1 x_1'] \\ & + \tfrac{\delta(1-\rho)}{2(1-\delta)(1-\delta\rho)}[B(x_1'', \theta_B) - c_1 x_1''] - \tfrac{\delta(1-\rho)}{2(1-\delta\rho)}[s_1(k_1 - x_1')] \\ & - \tfrac{\delta^2(1-\rho)^2}{2(1-\delta)(1-\delta\rho)}[s_1(k_1 - x_1'')] - \; -\tfrac{1}{2} s_1 x_1' - \tfrac{1}{2} s_1 k_1 - f_1 k_1\end{aligned}$$

The first term on the RHS of the payoff expression is gross benefit less production cost in the A states. The second term is gross benefit less production cost in B states that do not follow initial A state. The third term is gross benefit less production cost in B states that follow an initial A state. The fourth term is startup cost for a transition between an initial B state to the first A state. The fifth term is startup cost for all transitions from B states to A states after the first A state occurs. The sixth and seventh terms are startup costs for initial states B and A, respectively. The final term is investment cost. $\tilde{W}$ is quadratic and concave in its three arguments. The necessary conditions for an interior solution (with $x_1'' < k_1$) yield the following solution:

$$k_1^* = \theta_A - c_1 - 2(1-\delta)f_1 - (1-\delta\rho)s_1 \tag{24}$$

$$x_1''^* = \theta_B - c_1 + \delta(1-\rho)s_1 \tag{25}$$

$$x_1'^* = \theta_B - c_1 - (1-\delta)s_1 \tag{26}$$

$x_1'^*$ is an interim output level, occurring only during a sequence of periods following an initial draw of θ_B. Once state A occurs, output follows a Markov process that alternates between levels k_1^* and $x_1''^*$, with ρ equal to the probability that the same output persists into the next period. The expressions for k_1^* and $x_1''^*$, coupled with the demand function, yield the steady state prices given by (9) and (10).

Suppose there is an equilibrium in which technology 2 firms invest, startup whenever demand is in state A, and shut-down in transitions from A to B. Let $\tilde{p}_A$ be the equilibrium price for state A; the price(s) in state B is irrelevant for such a firm's profits. Let π_2^A be the firm's value of profits beginning in state A with active capacity and let π_2^B be the firm's value of profits beginning in state B with inactive capacity. These values satisfy the following two conditions.

$$\pi_2^A = \tilde{p}_A - c_2 + \delta\rho\pi_2^A + \delta(1-\rho)\pi_2^B \tag{27}$$

$$\pi_2^B = \delta\rho\pi_2^B + \delta(1-\rho)(\pi_2^A - s_2) \tag{28}$$

The value of initial investment for this technology 2 firm is:

$$v_2^e = \frac{1}{2}[\pi_2^A - s_2] + \frac{1}{2}\pi_2^B - f_2 \tag{29}$$

Solving for π_2^A and π_2^B in (27) and (28) and substituting on the RHS of (29) yields:

$$v_2^e = \frac{1}{2(1-\delta)}\tilde{p}_A - \frac{1}{2(1-\delta)}[c_2 + 2(1-\delta)f_2 + (1-\rho\delta)s_2] \tag{30}$$

The first term on the RHS of (30) is the expected DPV of revenue and the second term is the expected DPV of total costs of investment, startups and operation. Note however that under the assumption that $\Delta - (1-\rho\delta)(s_1 - s_2) > 0$, if a technology 1 firm made the same investment, startup and production decisions as this technology 2 firm, its expected DPV of revenue would be the same and its expected DPV of total cost would be lower. By Corollary 3.4 firms earn zero expected profit in equilibrium, so technology 1 firms cannot earn strictly greater profit from investing in stage 1 than technology 2 firms. This contradicts the supposition that technology 2 firms invest and produce in state A in equilibrium.

Part Two: Assume that, $\Delta - (1-\rho\delta)(s_1 - s_2) < 0$. If a technology 2 firm invests, starts whenever demand is in state A, and shuts down in transitions from A to B then the value of its profit is given by v_2^e in (30), where $\tilde{p}_A$ is the equilibrium price in state A. If a technology 1 firm invests, startups whenever demand is in state A, and shut-downs in transitions from A to B then the value of its profit is strictly less than v_2^e, given the assumption that $\Delta -$

$(1-\rho\delta)(s_1-s_2)<0$. Therefore, an equilibrium must involve technology 2 firms playing the role of 'swing' producers who startup in state A and shut-down in transitions from A to B. Furthermore, our assumption that $\Delta>0$ insures that technology 1 firms are the firms that operate in all demand states, rather than technology 2 firms.

The structure of equilibrium is fairly simple. Output is equal to total capacity, k_1+k_2, in all periods with high demand (state A). If the initial demand realization is low (state B) then startup and output are equal to x_1'; at the first transition to state A there is startup of technology 1 firms equal to k_1-x_1', startup of technology 2 firms equal to k_2, and output rises to k_1+k_2. In any transition from state A to B all technology 2 firms shut-down. In subsequent transitions from B to A all technology 2 firms startup. The payoff $\hat{W}$ to the planner may be expressed as a function of (x_1',k_1,k_2).

$$\begin{aligned}\hat{W}(x_1',k_1,k_2)=&\tfrac{1}{2(1-\delta)}[B(k_1+k_2,\theta_A)-c_1k_1-c_2k_2]+\tfrac{1}{2(1-\delta\rho)}[B(x_1',\theta_B)-c_1x_1']\\&+\tfrac{\delta(1-\rho)}{2(1-\delta)(1-\delta\rho)}[B(k_1,\theta_B)-c_1k_1]-\tfrac{\delta(1-\rho)}{2(1-\delta\rho)}[s_1(k_1-x_1')+s_2k_2]\\&-\tfrac{\delta^2(1-\rho)^2}{2(1-\delta)(1-\delta\rho)}[s_2k_2]-\ -\tfrac{1}{2}s_1x_1'-\tfrac{1}{2}[s_1k_1+s_2k_2]-f_1k_1-f_2k_2\end{aligned}$$

$\hat{W}(\cdot)$ is quadratic and concave. The necessary conditions for optimality yield the following welfare-maximizing expressions for x_1', k_1, and k_2:

$$\hat{x}_1'=\theta_B-c_1-(1-\delta)s_1 \tag{31}$$

$$\hat{k}_1=\theta_B-c_1+\delta(1-\rho)s_1+\frac{1-\rho\delta}{\delta(1-\rho)}[\Delta-(1-\rho\delta)(s_1-s_2)] \tag{32}$$

$$\hat{k}_2=\theta_A-c_2-(1-\delta\rho)s_2-2(1-\delta)f_2-\hat{k}_1 \tag{33}$$

The expressions for $\hat{k}_1$ and $\hat{k}_2$ may be used to derive steady state equilibrium prices in (11) and (12).

B Parameter Calibration

The cost parameters for technologies are built up from data on cost and technology fundamentals. In this section, we describe how marginal costs, investment costs, and startup costs are calibrated. Rather than using the average characteristics of existing plants, we calibrate the parameters using the characteristics for the latest generation of newly constructed generators. Using state-of-the-art technology parameters corresponds best to the notion of long run equilibrium in the model. It reflects what can expected as old facilities are phased out and are replaced by newer construction.

B.1 Marginal costs

We use a natural gas price of $3.50/MMBtu. This is roughly the average projected Henry Hub price (in constant 2019 dollars) over the next 30 years, as estimated by EIA [2019]. The natural gas price and heat rate are used to construct the marginal cost of production for the two types of natural gas generators. The marginal cost of a generator is computed as the product of its heat rate and fuel cost plus any variable maintenance costs and emissions costs.

To calculate generation costs we need information on the the conversion efficiency, or heat rate, of the generator as well as the cost of fuel used for production. We use data from the Annual Energy Outlook published by Energy Information Administration on the characteristics of new gas and coal fired power plants; EIA [2020a]. In particular, we use the heat rate, variable operating costs, and emission rate for natural gas combined cycle and conventional combustion turbine (Peaker) units.

The three cost components are shown in equation (34),

$$MC_j = \underbrace{HR_j * FC_j}_{\text{Fuel Cost}} + VOM_j + \underbrace{ER_j P_j}_{\text{Emission Cost}} \tag{34}$$

where,

HR = Heat Rate,

FC = Fuel Cost,

VOM = Variable Operating and Maintenance Cost,

ER = CO_2 Emission Rate,

P = CO_2 Price (Tax Rate).

Of these three, fuel costs have always dominated the marginal cost calculation. The fuel component is the raw cost of fuel multiplied by the conversion efficiency, or heat rate, of the generator. As shown in Table 5, a state-of-the-art GCC generator with a heat rate of 6.43 MWh/mmBTU facing a natural gas price of $3.50/mmBTU incurs fuel costs of $22.51 to generate a single MWh of electricity.

In addition to fuel costs, thermal plants emit potentially regulated pollutants such as SO2, NOx, and CO2. Creating these pollutants as a byproduct of combustion creates a cost for the firm in terms of operation of pollution control technologies and/or purchasing pollution permits. Generators in ERCOT have not had to pay for marginal emissions of SO2 of NOx in recent years, so we do not include expenses for these in emissions costs. Generators do not have to pay for CO2 emissions currently, so we do not include CO2 emissions costs in total marginal costs in our baseline simulations. For our carbon price policy simulations, we adjust marginal cost according to equation (34).

The cost of operating pollution control technologies are included in the Variable Operating and Maintenance (VOM) costs shown. VOM also includes any other costs that related to the amount of electricity produced. Costs from VOM account for roughly 5-20% of operating costs depending on the technology.

B.2 Investment Costs

For investment costs we include the overnight construction costs and annual fixed operations and maintenance costs reported in EIA [2020a]. We assume a 30 year lifetime for all three generator types. At the end of the lifetime of a generator, a firm could potentially decide to no longer participate in the market or not to reinvest in the same asset. However, since the distribution of demand is stationary and the technology parameters are unchanging, the decision problem 30 years in the future looks exactly the same as in the first period. Firms will always decide to reinvest in the same manner. Thus, we can model the decision to invest as a commitment to invest now and to reinvest in the same technology at the end of the life of the asset. The relevant investment cost for a firm at the initial investment stage is the present discounted construction costs and annual fixed OM costs over an infinite horizon. This is shown in equation (35),

$$IHC = \sum_{t=1}^{\infty}[(\tilde{\delta}^L)^{t-1}CC + \tilde{\delta}^t F] \tag{35}$$

where $\tilde{\delta}$ is the annual discount factor, L is plant lifetime, CC is plant construction cost, and F is annual O&M cost. We use annual discount rate of 5%, so $\tilde{\delta} = 1/1.05$. The reinvest part of the investment costs is small relative to initial costs the due to discounting over the long lifetime of the assets. Table 6 shows the infinite horizon investment costs. Note that wind and solar units which have the lowest marginal costs, have the highest investment costs. Gas turbines (Peakers), which have the highest marginal costs, are the least expensive technology to build. The tradeoff between low operating costs and high investment costs will lead to a mix of technologies even in the absence of startup costs.

The calibrated parameters for both investment costs and marginal costs are reasonable approximations of real costs. To avoid the perception of precision, we round marginal costs and investment costs to the nearest $1 and $5, respectively, for our analysis and the numbers reported in the paper.

B.3 Dynamic Parameters: Startup Costs and Minimum Output

Dynamics enter the model through two key parameters: startup costs and minimum output rates. Without either of these features, the model is completely static. Without startup costs, current actions have no implications for future profits. On the other hand, if generators are completely flexible in their output level, then they can avoid any startup costs by producing minuscule quantities.

Generator startup costs are a key feature of our application. Some economic studies of wholesale electricity markets have taken the position that startup costs are small enough that they can be safely ignored in an analysis of energy supply decisions.[24] This is likely true if one focuses on fuel and other energy costs associated with startups. However, the bulk of the opportunity cost of a generator startup is associated with additional maintenance and wear-and-tear on generators.[25] Kumar et al. [2012] estimate that capital and maintenance expenses comprise 80-98 percent of total startup costs, depending on generator and fuel type. Our startup cost figure for a GCC unit is based on Kumar et al. [2012], using their lower bound estimates for the capital and maintenance portion of startup costs as shown in Table 6. Our startup cost figure is broadly consistent with estimates from structural models using data from Spain in Reguant [2014]. Cullen [2015] reports structural estimates of startup costs using ERCOT data. He finds significantly higher startup costs for GCC units than does Reguant [2014]. We use the lower startup cost figure to be conservative in our simulations.

Minimum output rates are also an important feature of the model. If generators can ramp output down to arbitrarily low levels, then they will never have to incur startup costs. Higher minimum output rates sharpen the effect of start ups, while lower minimum output rates weaken the effect. A generator with lower minimum output is more flexible than a generator with a high minimum output. We calibrate the minimum output rate for gas combined cycle units using data from Cullen [2015]. As noted in the body of the paper, we assume that combustion gas turbines are completely flexible. That is, they have no startup costs or equivalently they have no minimum output. Either assumption eliminates any dynamic implications of their production. Gas turbines do incur small startup costs and have non-zero minimum output constraints. However, we assume that the startup costs and minimum output constraints are small enough for Peaker units as to be negligible.

[24] See for example, Borenstein et al. [2002], p. 1391.

[25] Perez-Arriaga and Batlle [2012] emphasize this point in their analysis of the effects of renewable intermittency on conventional power plant operations.

B.4 Wind and Solar Generation and Demand

Wind and solar generation and demand are assumed to evolve stochastically over time. Data for hourly wind turbine generation and installed capacity for 2019 are in the ERCOT Hourly Aggregated Wind Output file.[26] Wind generation capacity varies from 22,607 MW to 27,040 MW over the course of 2019. We divide hourly wind generation by installed wind capacity for each hour to arrive at hourly wind capacity factors for 2019. Hourly utility-scale solar PV generation for ERCOT for 2019 are drawn from the 2019 Fuel Mix Report.[27] Installed utility scale solar PV capacity ranges from 1,719 MW at the start of 2019 to 2,281 MW at end of year.[28] We divide hourly solar PV generation by installed capacity to arrive at hourly solar capacity factors. There is also considerable distributed solar PV generation capacity at residential and business sites in ERCOT. We do not have data on total distributed solar generation since this solar generation is behind the customer meter. Hourly load data is net of distributed solar generation. Hourly load data for ERCOT for 2019 appears in the Hourly Load Data Archives.[29]

We use hourly data on wind and solar capacity factors and load (demand quantity) for ERCOT from 2019 to estimate a forecasting model for wind, solar, and load. In order to conform to the Markovian structure of the planner's DP problem, our forecasting model includes lagged load and lagged wind and solar capacity factors as dependent variables. We estimate the following regression equation for hourly load:

$$L_t = \sum_{h=1}^{24} (\beta_{1h} D_{ht} + \beta_{2h} D_{ht} L_{t-1} + \beta_{3h} D_{ht} L_{t-1}^2 + \beta_{4h} D_{ht} \nu_{t-1} + \beta_{5h} D_{ht} \sigma_{t-1}) + \epsilon_t^L \tag{36}$$

where L_t is net load in period t, D_{ht} is the hour-of-day h dummy for period t, the β terms are coefficients, and ϵ_t^L is the one-hour-ahead prediction error. The adjusted R^2 is very high at 0.991.

[26]This file may be found at http://www.ercot.com/gridinfo/generation
[27]This file is from http://www.ercot.com/gridinfo/generation
[28]ERCOT Fact Sheets for 2019 and 2020
[29]See http://www.ercot.com/gridinfo/load/load_hist

We estimate the following two regression equations for wind capacity factor and solar capacity factor:

$$\nu_t = \sum_{h=1}^{24} (\alpha_{1h} D_{ht} + \alpha_{2h} D_{ht} \nu_{t-1} + \alpha_{3h} D_{ht} \nu_{t-1}^2 + \alpha_{4h} D_{ht} L_{t-1} + \alpha_{5h} D_{ht} \sigma_{t-1}) + \epsilon_t^\nu \tag{37}$$

$$\sigma_t = \sum_{h=1}^{24} (\phi_{1h} D_{ht} + \phi_{2h} D_{ht} \sigma_{t-1} + \phi_{3h} D_{ht} \sigma_{t-1}^2 + \phi_{4h} D_{ht} L_{t-1} + \phi_{5h} D_{ht} \nu_{t-1}) + \epsilon_t^\sigma \tag{38}$$

The prediction errors are ϵ_t^ν and ϵ_t^σ for wind and solar, respectively. The adjusted R^2 values are 0.971 and 0.975 for the wind and solar regression equations, respectively. Hour ahead load, wind, and solar may be predicted with a high degree of accuracy, though predicted levels farther in the future are considerably less accurate.

The demand vertical intercept ψ is used as a state variable in the stage two DP problem. Load quantities are translated to vertical intercepts using inverse demand equation (17) and the average wholesale electricity price $\bar{p}$ observed in the data. A load quantity L_t corresponds to vertical intercept $\psi_t = \bar{p} + bL_t$.

C Computation

Computing the long run competitive equilibrium for the dynamic model entails solving a two stage problem of surplus maximization. We use backward recursion. Stage 2 is an infinite horizon stochastic dynamic programming problem that is conditional on a vector of generation capacities. Stage 1 involves choosing a vector of generation capacity amounts to maximize stage 2 expected surplus less capacity investment cost. The state vector for the stage 2 DP problem is $(x, \boldsymbol{\theta}) = (x, \psi, \nu, \sigma, h)$. State components x, ψ, ν, and σ are continuous variables. The vector $\boldsymbol{\theta}$ constitutes the information set for constructing beliefs about demand shocks and renewable capacity factors in future periods. The model for forecasting future demand and renewable capacity factors is described in Section 4.2 and Appendix B.4.

We use the nonlinear certainty equivalent (NLCEQ) method described in Cai et al. [2017] to solve the stage 2 DP problem. This method approximates the optimal policy function at a finite collection of state values by solving deterministic finite horizon optimization problems in which future values of exogenous random state variables are set equal to their expected values conditional on the current state. This method can accommodate large DP problems with continuous state variables and control variables with inequality constraints. Given our assumptions regarding demand and costs, each finite horizon optimization problem is a quadratic programming problem with linear constraints. Each quadratic programming problem solves quickly. This is important, since our approach requires solving 8,760 (# hours/year) of these problems in stage 2 for each potential stage 1 capacity vector.

A complication for solving the stage 2 DP problem is that the discount factor δ for hourly time periods is very close to one. This feature of the problem would potentially require a long time horizon T to adequately capture the effect of future periods on the policy function using the NLCEQ method.[30] It turns out that the policy function is not very sensitive to δ for our problem and we are able to capture the future value of startup and shut-down decisions using $T = 18$ hourly periods for the quadratic programming problem.

We use the computed policy function to calculate one year's worth of total surplus at a sequence of optimal x values and realized values of demand and renewable capacity factors. The total surplus payoff for stage 2 is used as input to solve for optimal stage 1 capacity investments. By Proposition 3.1 the surplus maximizing allocation corresponds to a competitive equilibrium allocation.

[30] An alternative approach is to discretize the state space and use policy function iteration. This is the approach used in Cullen and Reynolds [2017]. Policy function iteration solves quickly even with δ close to one. However, the discretization required for this approach must be fairly coarse given the large number of state variables for our problem, and a coarse discretization introduces errors into policy function computation.

References

Adlakha, S., Johari, R., and Weintraub, G. (2015). Equilibria of dynamic games with many players. *Journal of Economic Theory*, 156:269–316.

Bigerna, S. and Polinori, P. (2014). Electricity demand in wholesale Italian market. *The Energy Journal*, 35(3).

Blanford, G., Merrick, J., and Young, D. (2014). A clean energy standard analysis with the US-REGEN model. *The Energy Journal*, 35(SI1).

Borenstein, S., Bushnell, J., and Wolak, F. (2002). Measuring market inefficiencies in California's restructured wholesale electricity market. *American Economic Review*, 92:1376–1405.

Bushnell, J. (2010). Building blocks: Investment in renewable and non-renewable technologies. In Moselle, B., Padilla, J., and Schmalensee, R., editors, *Harnessing Renewable Energy in Electric Power Systems*. RFF Press.

Bushnell, J. and Novan, K. (2018). Setting with the sun: The impacts of renewable energy on wholesale power markets. UC Davis Energy Economics Program Working Paper # 020.

Cai, Y., Judd, K., and Steinbuks, J. (2017). A nonlinear certainty equivalent approximation method for dynamic stochastic problems. *Quantitative Economics*, 8.

Cho, I.-K. and Meyn, S. P. (2011). Efficiency and marginal cost pricing in dynamic competitive markets with friction. *Theoretical Economics*, 5:215–239.

Chyong, C. K., Guo, B., and Newbery, D. (2020). The impact of a carbon tax on the co2 emissions reduction of wind. *The Energy Journal*, 41.

Cullen, J. (2013). Measuring the environmental benefits of wind-generated electricity. *American Economic Journal: Economic Policy*, 5(4):107–133.

Cullen, J. (2015). Dynamic response to environmental regulation in the electricity industry. Working Paper, Washington University.

Cullen, J. and Reynolds, S. (2017). Market dynamics and investment in the electricity sector. University of Arizona Working Paper # 17-05.

Deryugina, T., MacKay, A., and Reif, J. (2020). The long-run dynamics of electricity demand: Evidence from municipal aggregation. *American Economic Journal: Applied Economics*, 12(1):86–114.

EIA (2019). Annual Energy Outlook 2019. Technical report, U.S. Energy Information Administration.

EIA (2020a). Cost and performance characteristics of new generating technologies, annual energy outlook 2020. Technical report, U.S. Energy Information Administration.

EIA (2020b). Preliminary monthly electric generator inventory. Technical report, U.S. Energy Information Administration.

EIA (2021). U.S. wind energy production tax credit extended through 2021. Technical report, Today in Energy, U.S. Energy Information Administration.

Ericson, R. and Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies*, 62:53–82.

Fell, H. and Kaffine, D. T. (2014). A one-two punch: Joint effects of natural gas abundance and renewables on coal-fired power plants. Colorado School of Mines Working Paper 2014-10.

Fell, H. and Linn, J. (2013). Renewable electricity policies, heterogeneity, and cost effectiveness. *Journal of Environmental Economics and Management*, 66:688–707.

FERC (2014). Staff analysis of uplift in RTO and ISO markets. Technical report, Federal Energy Regulatory Commission.

Gowrisankaran, G., Reynolds, S., and Samano, M. (2016). Intermittency and the value of renewable energy. *Journal of Political Economy*, 124:1187–1234.

Green, R. and Vasilakos, N. (2011). The long-term impact of wind power on electricity prices and generating capacity. Conference Proceedings for IEEE Power and Energy Society General Meeting.

Hopenhayn, H. (1990). Industry equilibrium dynamics: A general competitive theory. Working Paper.

Hortacsu, A. and Puller, S. (2008). Understanding strategic bidding in multi-unit auctions: A case study of the Texas electricity spot market. *RAND Journal of Economics*, 39:86–114.

IPCC (2014). Climate change 2014: Mitigation of climate change. Technical report, Intergovernmental Panel on Climate Change 2014.

Johnsen, T. (2001). Demand, generation and price in the Norwegian market for electric power. *Energy Economics*, 23:227–251.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica*, 50:24–52.

Kaffine, D. T., McBee, B. J., and Lieskovsky, J. (2013). Emissions savings from wind power generation in Texas. *The Energy Journal*, 34(1):155–175.

Kumar, N., Besuner, P., Lefton, S., and Agan, D. (2012). Power plant cycling costs. Technical report. NREL Report SR-5500-55433.

LaRiviere, J. and Lu, X. (2020). Transmission constraints, intermittent renewables and welfare. *Environmental Science*.

Lucas, R. and Prescott, E. (1971). Investment under uncertainty. *Econometrica*, 39.

Novan, K. (2015). Valuing the wind: Renewable energy policies and air pollution avoided. *American Economic Journal: Economic Policy*, 8(1):291–326.

O'Neill, R., Sotkiewicz, P., Hobbs, B., Rothkopf, M., and Stewart, W. (2005). Efficient market-clearing prices in markets with nonconvexities. *European Journal of Operational Research*, 164:269–285.

Patrick, R. and Wolak, F. (2001). Estimating the customer-level demand for electricity under real-time market prices. NBER Working Paper No. 8213.

Perez-Arriaga, I. and Batlle, C. (2012). Impacts of intermittent renewables on electricity generation system operations. *Economics of Energy & Environmental Policy*, 1(2).

Reguant, M. (2014). Complementary bidding mechanisms and startup cost in electricity markets. *Review of Economic Studies*, 124:1708,1742.

Rob, R. (1991). Learning and capacity expansion under demand uncertainty. *Review of Economic Studies*, 58:655–675.

Sensfuß, F., Ragwitz, M., and Genoese, M. (2008). The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany. *Energy Policy*, 36.

Stokey, N. and Lucas, R. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

Troy, N., Denny, E., and O'Malley, M. (2010). Base-load cycling on a system with significant wind penetration. *IEEE Transactions on Power Systems*, 25(2):1088–1097.

Williamson, O. (1966). Peak-load pricing and optimal capacity under indivisibility constraints. *American Economic Review*, 61:810–827.

Wolfram, C. (1999). Measuring duopoly power in the British electricity spot market. *American Economic Review*, 89:805–826.

Xie, L., Puller, S., Ilic, M., and Oren, S. (2013). Quantifying benefits of demand response and look-ahead dispatch in systems with variable resources. PSERC Publication 13-41.

Tables

Table 1: **Electricity Generator Parameters**

		GCC	Peaker	Wind	Solar
Marginal Cost	($/MWh)	25	39	0	0
Min. Output	(% of Capacity)	70	0	0	0
Startup Cost	($/MW)	80	0	0	0
Investment Cost	($/kW)	1,690	1,065	2,180	1,910

Table 2: **Equilibrium Results with No Policy Incentives**

		Static Model	Dynamic Model
Capacity Investment (1,000s of MW)			
	GCC	48.4	45.4
	Peaker	6.5	6.2
	Wind	0	4.8
	Solar	5.5	9.1
Average Wholesale Electricity Price ($ per MWh)			
		34.41	35.44
Within Day Standard Deviation of Wholesale Electricity Price			
		10.2	11.7
Renewable Energy Penetration (as % of Total energy Generation)			
		3%	9%
Startup Costs (as % of Operating Cost)			
	GCC units	-	0.4%

Table 3: **Equilibrium Results with Policies to Achieve CO2 Reduction Targets**

CO2 Reduction Target	None	20%	40%	60%
Policy Parameters to Achieve Targets				
Renewable Investment Subsidy (% ITC)	-	9.2	16.3	32.0
Carbon Tax ($ per ton CO2)	-	7.60	14.43	31.55
Average Wholesale Electricity Price ($ per MWh)				
Renewable Subsidy Policy	35.44	36.01	34.57	31.21
Carbon Tax Policy	35.44	38.86	40.25	43.33
Within Day Standard Deviation of Electricity Price				
Renewable Subsidy Policy	11.7	12.3	15.0	17.2
Carbon Tax Policy	11.7	12.5	16.0	20.9
Wind Penetration (% of Total Generation from Wind Turbines)				
Renewable Subsidy Policy	3.9	19.4	36.5	53.3
Carbon Tax Policy	3.9	19.2	35.5	51.9
Solar Penetration (% of Total Generation from Solar PV)				
Renewable Subsidy Policy	4.9	8.1	10.5	14.0
Carbon Tax Policy	4.9	7.4	9.7	12.0
Policy Effectiveness ($ per ton CO2 reduction)				
Renewable Subsidy Policy	-	3.88	7.75	14.33
Carbon Tax Policy	-	3.81	7.35	12.27
Renewable Curtailment (% of Total Potential Generation)				
Renewable Subsidy Policy	-	-	0.1	3.5
Carbon Tax Policy	-	-	-	1.8
GCC Startup Cost (% of Operating Cost)				
Renewable Subsidy Policy	0.4	1.1	3.2	7.3
Carbon Tax Policy	0.4	1.1	3.2	7.5

Table 4: **Static and Dynamic Model Results for Renewable Investment Subsidies**

CO2 Reduction Target	None	20%	40%	60%
Policy Parameters to Achieve Targets (% ITC)				
Static Model	1.5	6.3	10.4	28.2
Dynamic Model	-	9.2	16.3	32.0
Average Wholesale Electricity Price ($ per MWh)				
Static Model	34.41	34.41	34.35	31.26
Dynamic Model	35.44	36.01	34.57	31.21
Within Day Standard Deviation of Electricity Price				
Static Model	10.2	10.3	10.5	13.7
Dynamic Model	11.7	12.3	15.0	17.2
Policy Effectiveness ($ per ton CO2 reduction)				
Static Model	-	2.00	3.94	8.41
Dynamic Model	-	3.88	7.75	14.33
Renewable Curtailment (% of Total Potential Generation)				
Static Model	-	-	-	2.6
Dynamic Model	-	-	0.1	3.5

Table 5: **Marginal Cost Parameters**

		GCC	Peaker
Heat Rate	(mmBTU/MWh)	6.43	9.90
Fuel Cost	($/mmBTU)	3.50	3.50
Marginal Fuel Cost	($/MWh)	22.51	34.65
CO2 Rate	(ton/MWh)	0.37	0.58
CO2 Cost	($/ton)	-	-
Emission Cost	($/MWh)	-	-
VO&M	($/MWh)	2.56	4.52
Total MC	($/MWh)	**25.07**	**39.17**

Table 6: **Investment and Startup Cost Parameters**

		GCC	Peaker	Wind	Solar PV
Construction Cost (CC)	($/kW)	1,082	709	1,268	1,232
Lifetime (L)	(years)	30	30	30	30
Fixed O&M Cost (F)	($/kW/yr)	14	7	26	15
Infinite Horizon Cost (IHC)	($/kW)	**1,691**	**1,065**	**2,180**	**1,910**
Startup Cost	($/MW)	80	-	-	-
Minimum Output	(% of Capacity)	70	-	-	-

Figure Captions

Figure 1. Investment Incentives
(Parameters: $\delta = 0.9$, $s_1 = 6\Delta$)

Figure 2. Hourly Average Data

Figure 3. Monthly Average Data

Figure 4. Hourly Residual Demand

Figure 5. Equilibrium GCC Capacity Investment (1,000s of MW)

Figure 6. Equilibrium Peaker Capacity Investment (1,000s of MW)

Figure 7. Hourly Prices with Investment Subsidy

Figure 8. Hourly Prices with Carbon Price

Figure 9. Equilibrium GCC Capacity Investment (1,000s of MW)

Figure 10. Equilibrium Peaker Capacity Investment (1,000s of MW)

Figures

Figure 1: Investment Incentives
(Parameters: $\delta = 0.9$, $s_1 = 6\Delta$)

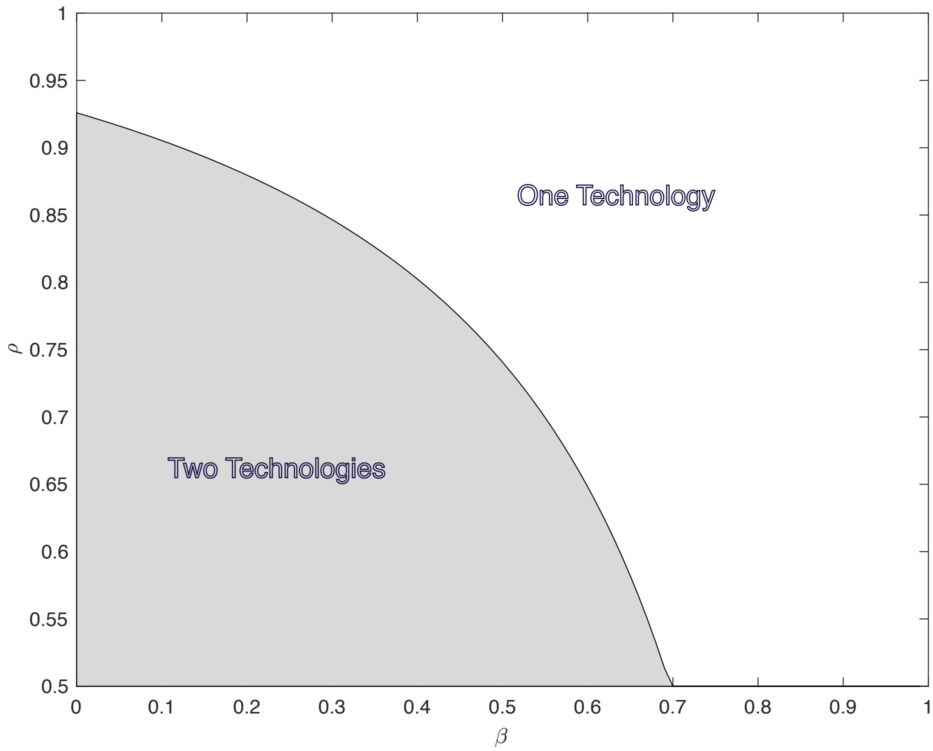

Figure 2: Hourly Average Data

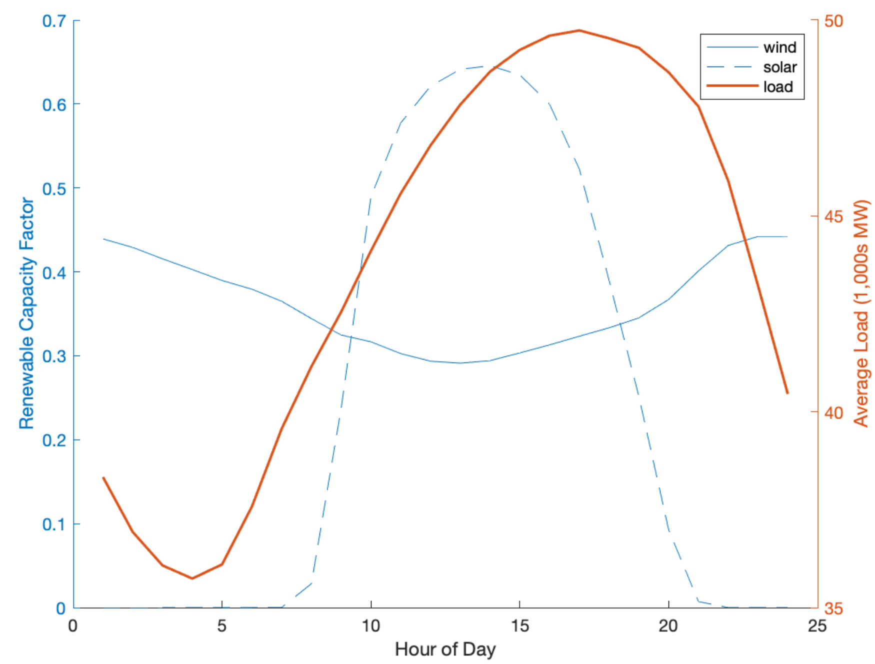

Figure 3: Monthly Average Data

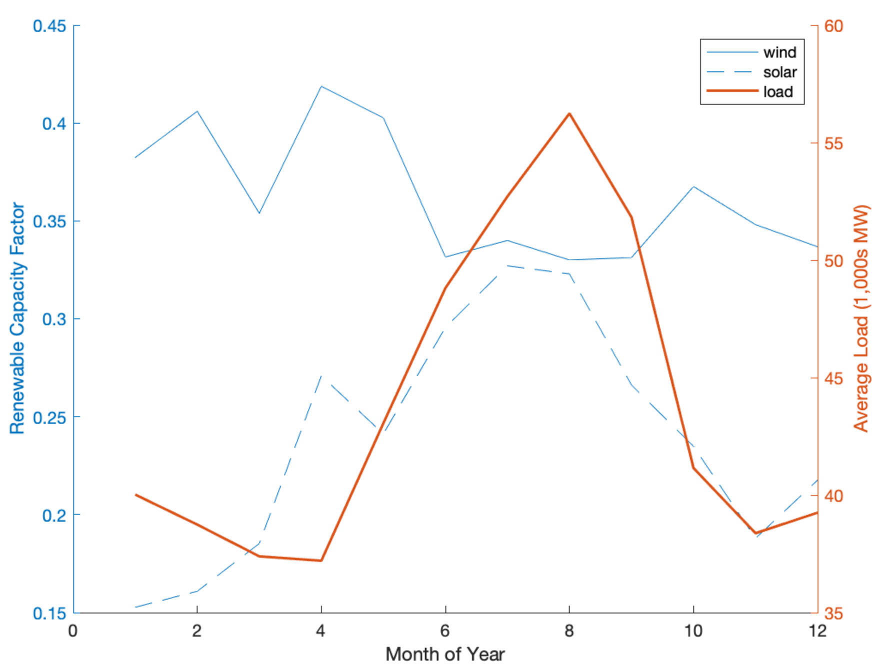

Figure 4: Hourly Residual Demand

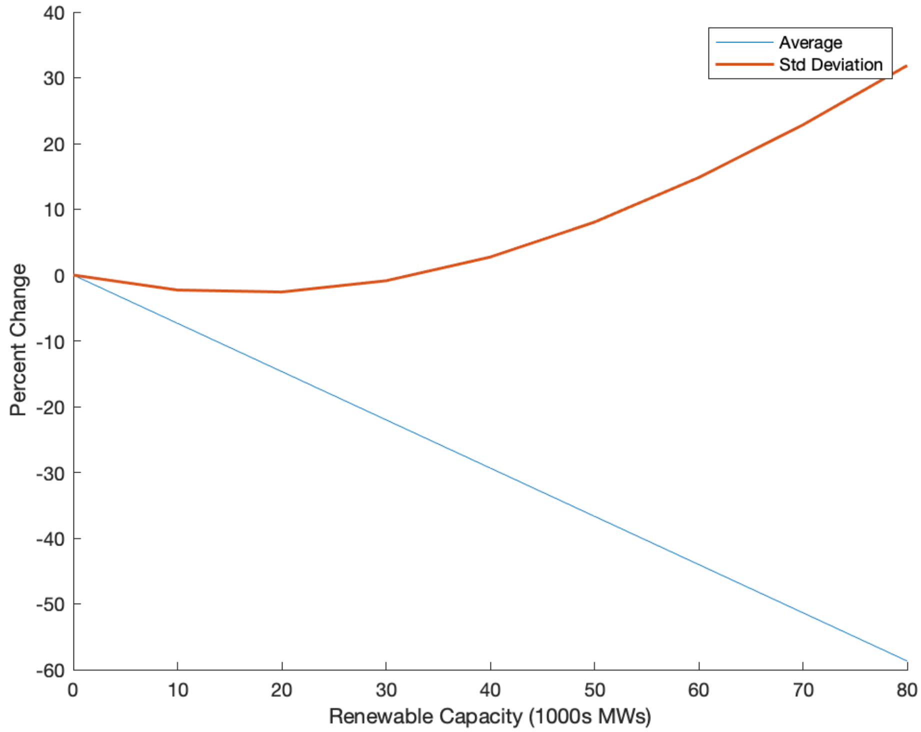

Figure 5: Equilibrium GCC Capacity Investment (1,000s of MW)

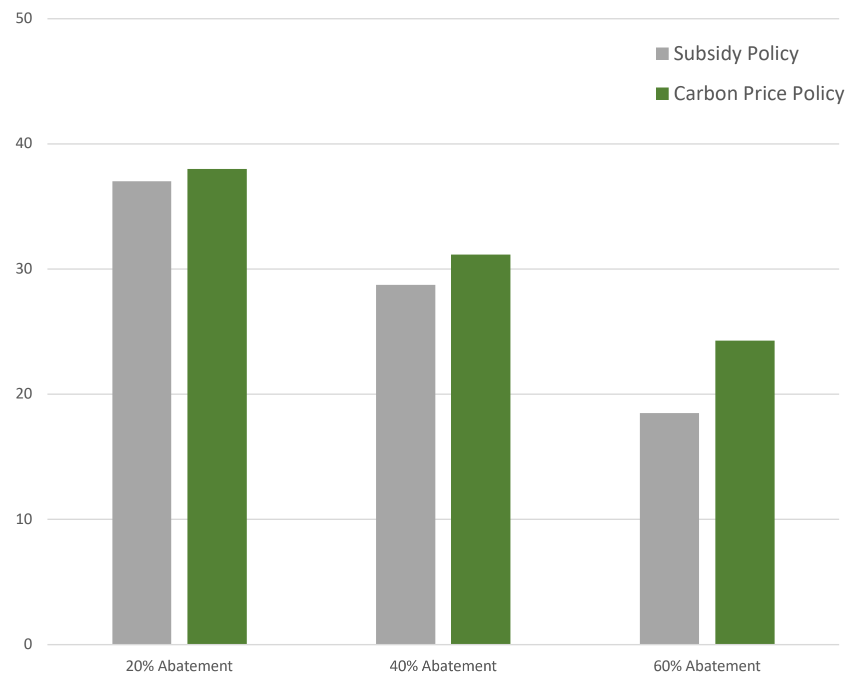

Figure 6: Equilibrium Peaker Capacity Investment (1,000s of MW)

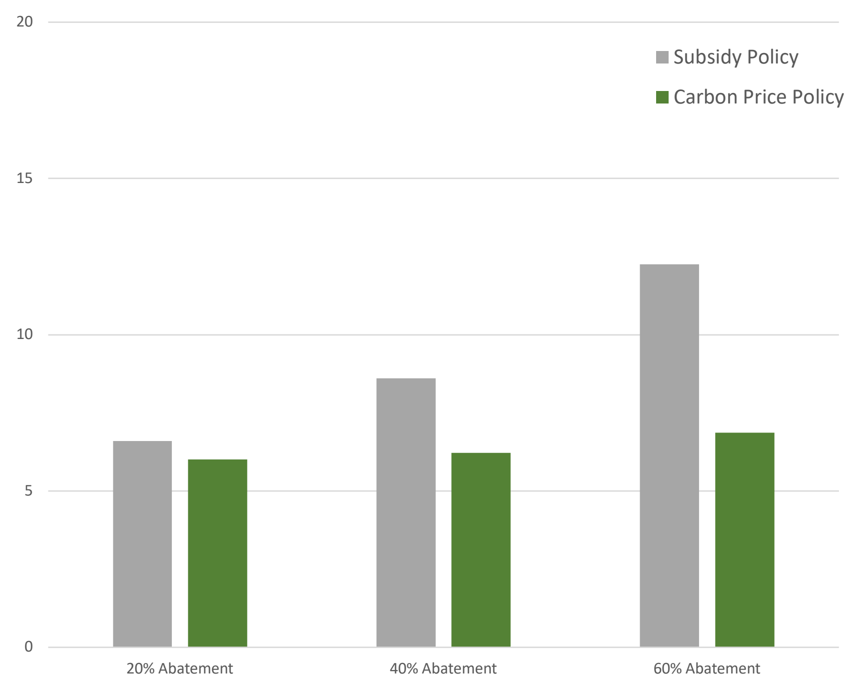

Figure 7: Hourly Prices with Investment Subsidy

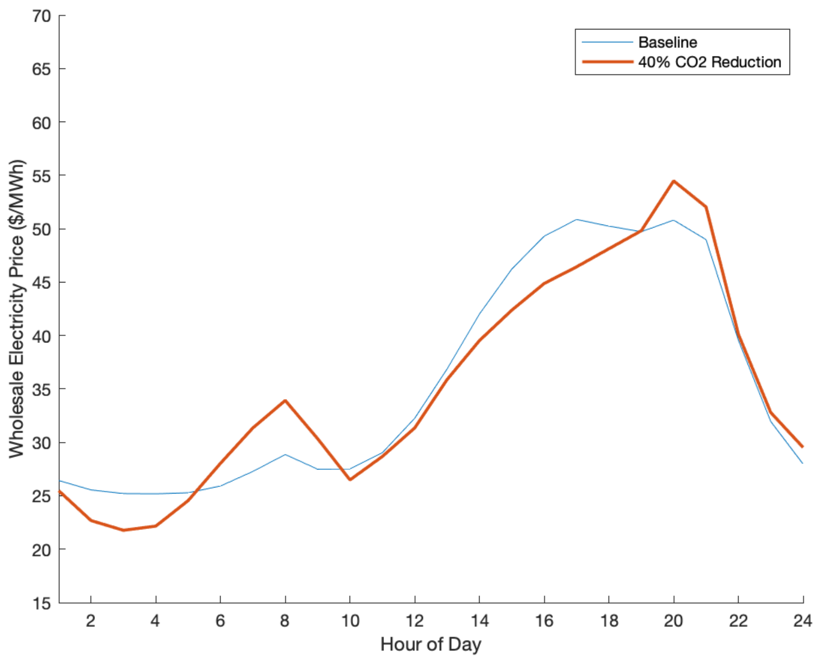

Figure 8: Hourly Prices with Carbon Price

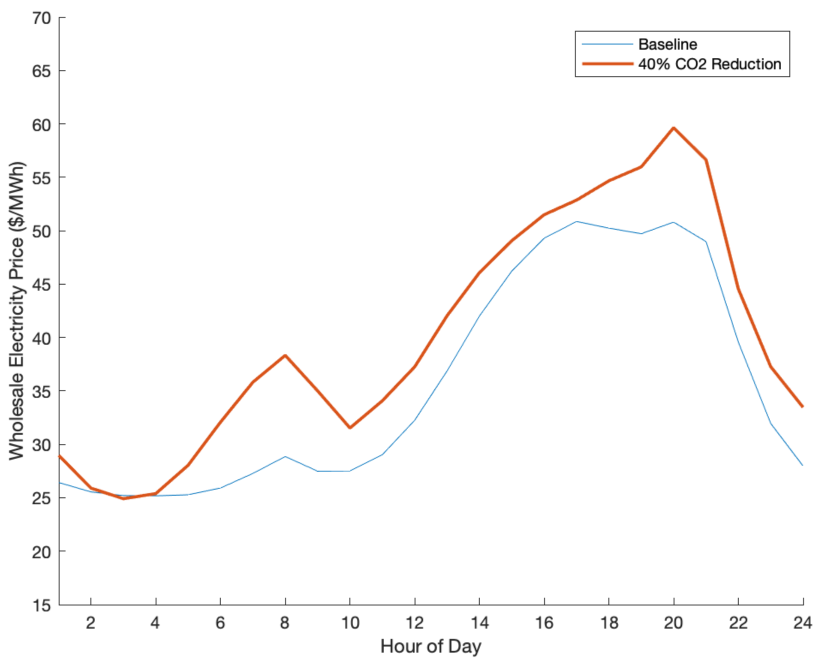

Figure 9: Equilibrium GCC Capacity Investment (1,000s of MW)

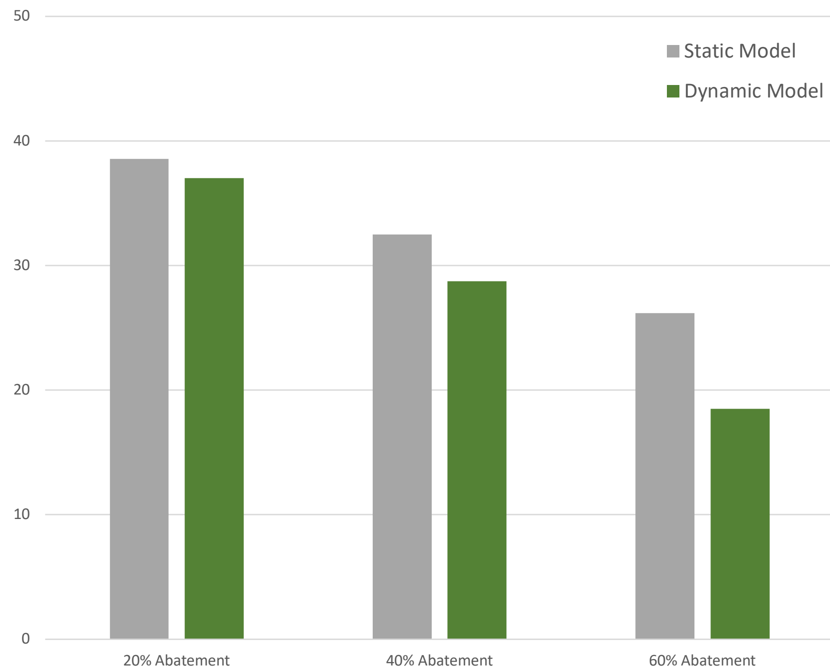

Figure 10: Equilibrium Peaker Capacity Investment (1,000s of MW)